

统计稀疏学习中的贝叶斯非参数建模方法及其应用研究

TONGJI XISHU XUEXIZHONG DE BEIYESI FEICANSHU JIANMO FANGFA JIQI YINGYONG YANJIU

何岩 著





浙江工商大学出版社



统计稀疏学习中的贝叶斯非参数
建模方法及其应用研究


ISBN 978-7-5178-0188-7

9 787517 801887 >
定价：18.00元

统计稀疏学习中的贝叶斯非参数建模方法及其应用研究

TONGJI XISHU XUEXIZHONG DE BEIYESI FEICANSHU JIANMO FANGFA JIQI YINGYONG YANJIU



何岩 著



浙江工商大学出版社

图书在版编目(CIP)数据

统计稀疏学习中的贝叶斯非参数建模方法及其应用研究/何岩著. —杭州：浙江工商大学出版社，2014.4
ISBN 978-7-5178-0188-7

Ⅰ.①稀… Ⅱ.①何… Ⅲ.①贝叶斯估计—数学模型—研究 Ⅳ.①O211.67

中国版本图书馆CIP数据核字(2013)第319128号

统计稀疏学习中的贝叶斯非参数建模方法及其应用研究
何 岩著

责任编辑 王玲娜 刘 韵
责任印制 包建辉
出版发行 浙江工商大学出版社
(杭州市教工路198号 邮政编码310012)
(E-mail:zjgsupress@163.com)
(网址:http://www.zjgsupress.com)
电话:0571-88904980,88831806(传真)
排 版 杭州朝曦图文设计有限公司
印 刷 杭州杭新印务有限公司
开 本 710mm×1000mm 1/16
印 张 6
字 数 114千
版印次 2014年4月第1版 2014年4月第1次印刷
书 号 ISBN 978-7-5178-0188-7
定 价 18.00元

浙江工商大学出版社营销部邮购电话 0571-88804228

目　录

CONTENTS

第 1 章
绪　　论

1.1 研究背景和意义

统计学习(Statistical Learning)是一种专门研究小样本情况下机器学习规律的理论,在这种体系下的统计推理规则不仅考虑了对渐近性能的要求,而且追求在现有有限信息的条件下得到最优结果。近年来,统计学习领域的学者结合稀疏特性对传统统计学习理论和方法进行了丰富和拓展,基于稀疏的统计学习逐步成为统计学习与信息处理的重要研究方向,其在数据挖掘、内容检索、基因数据分析等诸多领域得到了广泛应用。

1.1.1 稀疏编码的生物感知基础

对于稀疏的研究,最早源于对神经科学和脑科学认知的研究成果。1954 年,Attneave 最先提出视觉感知的目标就是产生一个外部输入信号的有效表示。Barlow 在 1961 年基于信息论提出了"有效编码假设",认为初级视觉皮层神经细胞的主要功能是去除输入刺激的统计相关性。20 世纪 60 年代末期,神经生理研究已表明了初级视觉皮层下细胞的感受野具有显著的方向敏感性,单个神经元仅对某一频段的信息呈现较强的反映,如特定方向的边缘、线段、条纹等图像特征,其空间感受野被描述为具有局部性、方向性和带通特性的信号编码滤波器,而每个神经元对这些刺激的表达则采用了稀疏编码(Sparse Coding)原则,将图像在边缘、端点、条纹等方面的特性以稀疏编码的形式进行描述。1996 年,Olshausen 和 Field 在 *Nature* 上发表论文,指出自然图像经过稀疏编码后得到的基函数类似于 V1 区内简单细胞感受野的反应特性。这种稀疏编码模型提取的基函数首次成功模拟了 V1 区内简单细胞感受野的三个响应特性:空间域的局部性、时域和频域的方向性和选择性。考虑到基函数的过完备性(基函数维数大于输出神经元的个数),

Olshausen 和 Field 在 1997 年提出了一种超完备基的稀疏编码算法，利用基函数和系数的概率密度模型成功地模拟了 V1 区简单细胞感受野。

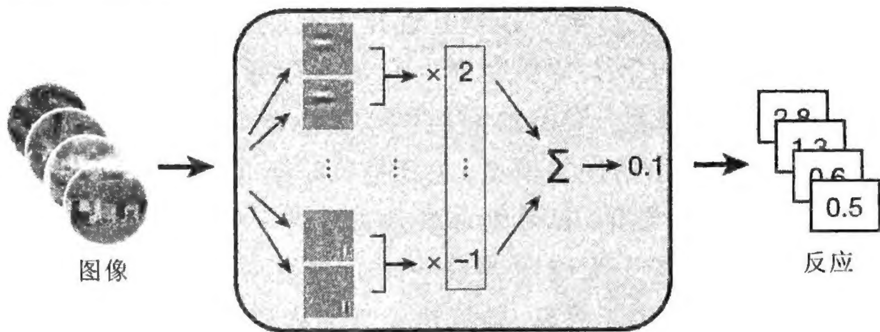
近年来，人们从神经生物学机理模型和计算机科学可计算模型等角度对稀疏编码理论进行了广泛的研究，并对生物视觉、脑科学的发展产生了重要的影响。Kay K. N. 和 Gallant 等从神经生理学机制上揭示了稀疏表达作为一种广泛的视觉先验，精确地定位于人类大脑视觉皮层多个功能区（如 V1、V2 区），并在视觉认知和推理过程中发挥着重要作用，例如图 1.1 显示 Kay K. N. 等人对图像识别的建模过程。这个过程的第一阶段是模型估计，即对每个测试者观看一组自然图像时产生的功能磁共振成像（FMRI）数据进行记录，再根据这些数据为每类图像构建一个定量的感受野模型，称为相对感受野模型（receptive-field model）。该模型基于 Gabor 滤波金字塔，并依照细胞感受野的三个特性进行描述。第二个阶段是图像识别，让每个测试者观看另外一组与先前测试图片不同的自然影像，并记录当时的功能磁共振成像数据。然后通过第一阶段构建的相对感受野模型来计算这组自然图像，预测每一张图片的功能磁共振成像数据，将预测数据与实际测量数据相对比，选取最相近的预测数据，从而得到测试者观看的图片。这些研究强调人类的认知和推理过程，不仅需要依据完整的信息输入，更需要依据视觉输入中的很少一部分典型特征，即依据某种稀疏编码求解，这为解决视觉认知问题提供了重要的生理学模型借鉴。

1.1.2 稀疏编码的信号表达基础

稀疏作为一种重要的数据编码与表达方式，不仅在人类的视觉认知机理上具有明确的理论依据，而且在信号表达与重建的理论方面得到了严格证明和推导。Donoho，Tao，Candès 和 Baraniuk 等提出的压缩感知（Compressive Sensing，CS）理论，从信号表达的角度证明了稀疏表达是高维信号（比如音频、视频等）在特定基向量（比如傅里叶基、小波基等）或“字典”上的一种自然表达。可压缩信号的少量随机线性投影即包含了重构和处理的足够信息，利用信号可压缩的先验知识和少量全局的线性测量可以获得精确的信号重建。在压缩感知理论基础上发展的约束优化求解策略为信号的稀疏表达提供了近似最优的可计算模型。同时，通过学习生成的自适应过完备冗余字典对稀疏表示求解的促进作用，引发研究者对字典学习算法的大量研究。2008 年 Candès 证明了如果随机正交模型条件成立，则能够以高概率恢复稀疏矩阵，从而从理论上证明矩阵填充（Matrix Completion）的可解性。基于上述理论的证明，目前统计稀疏学习已经广泛应用在信号压缩、图像处理、模式识别、机器学习等领域。

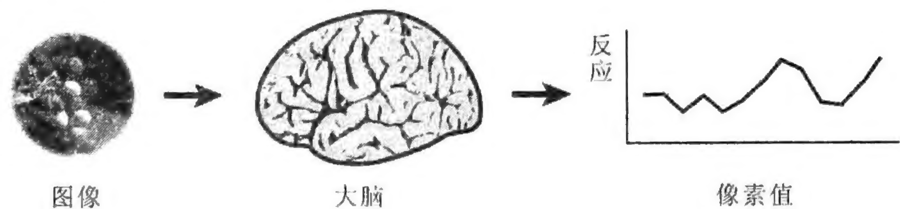
第一步,模型估计

为图像的每个像素估计一个感受野模型。

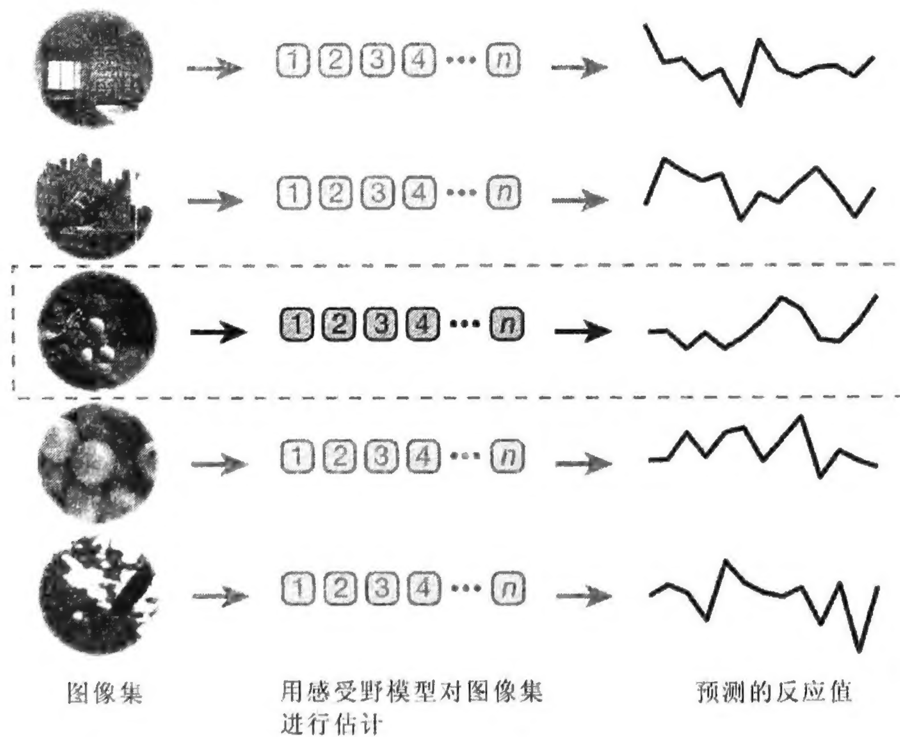


第二步:定义图像

(1) 每个图像度量大脑的反应。



(2) 用感受野模型为图像集预测大脑反应。



(3) 选择与测量的大脑反应值最接近的预测值进行标记。

图 1.1 Kay K. N. 对图像识别的建模过程

1.1.3 贝叶斯非参数方法

贝叶斯学习理论用于统计学习领域是近几年发展起来的最重要的主流研究方向，是目前JMLR，NIPS，ICML等机器学习领域国际重要期刊与会议论文的热点讨论内容。贝叶斯学习理论将先验知识与样本信息相结合、依赖关系与概率表示相结合，是不确定知识表示的理想模型，尤其是贝叶斯非参数方法所表现的灵活性引起研究者的广泛关注。然而，贝叶斯非参数方法并不是新的方法，早在1973年，Ferguson就提出了以带有无限维度参数空间的参数模型来表示先验的贝叶斯非参数方法。但由于推理方法不成熟、计算机运算速度慢等原因，贝叶斯非参数方法一直停留于理论研究。近年来，高速计算机的快速发展解决了边缘概率积分的复杂计算问题，同时，MCMC方法、EM算法，以及关于边缘概率计算的近似算法如变元推理等计算方法的发展，大大扩展了贝叶斯非参数方法的应用领域。2001年，Tipping提出了"稀疏贝叶斯学习(Sparse Bayesian Learning)"的概念，利用层次先验、核函数和贝叶斯推断，给出了稀疏的相关向量机的学习方法，从而为统计稀疏学习方法提供了新的研究思路。此后，基于贝叶斯非参数的统计稀疏学习方法层出不穷，并在文本内容检索、基因数据分析、计算机视觉等领域获得应用。结合贝叶斯非参数方法的不确定性知识表达形式，综合先验知识的增量学习特性和非参数的模型灵活性，研究基于贝叶斯非参数的统计稀疏学习方法的独特性能和技术优势，并在应用中对其模型、方法和算法性能进行全面评估至关重要。

1.1.4 统计稀疏学习方法的视觉应用

在各种应用研究中，视觉任务面临的往往是有噪声、高维、大批量及多样性的数据样本，而且需要对数据内容进行高层次、结构性的语义分析和自动注解，这对当前的统计学习方法提出了很大的挑战。基于稀疏表达的视觉应用近年来已经取得了一些研究成果，比如：人脸识别、图像超分辨率、图像降噪、背景建模、运动分割等应用。基于贝叶斯非参数的统计稀疏学习方法的视觉应用还处于起步阶段，但应用的效果却让人印象深刻，例如，图1.2是基于贝叶斯方法学习的分类字典，图1.3是图像分割效果，图1.4是基于贝叶斯非参数方法得到的图像插值效果。研究统计稀疏学习中的贝叶斯非参数方法在视觉任务中的应用对于全面评估相应方法、模型和算法性能至关重要，也有助于深入理解贝叶斯非参数统计稀疏学习方法的理论价值，并为结合稀疏表达的贝叶斯非参数统计学习方法的有效性提供了很好的验证平台和应用示例。

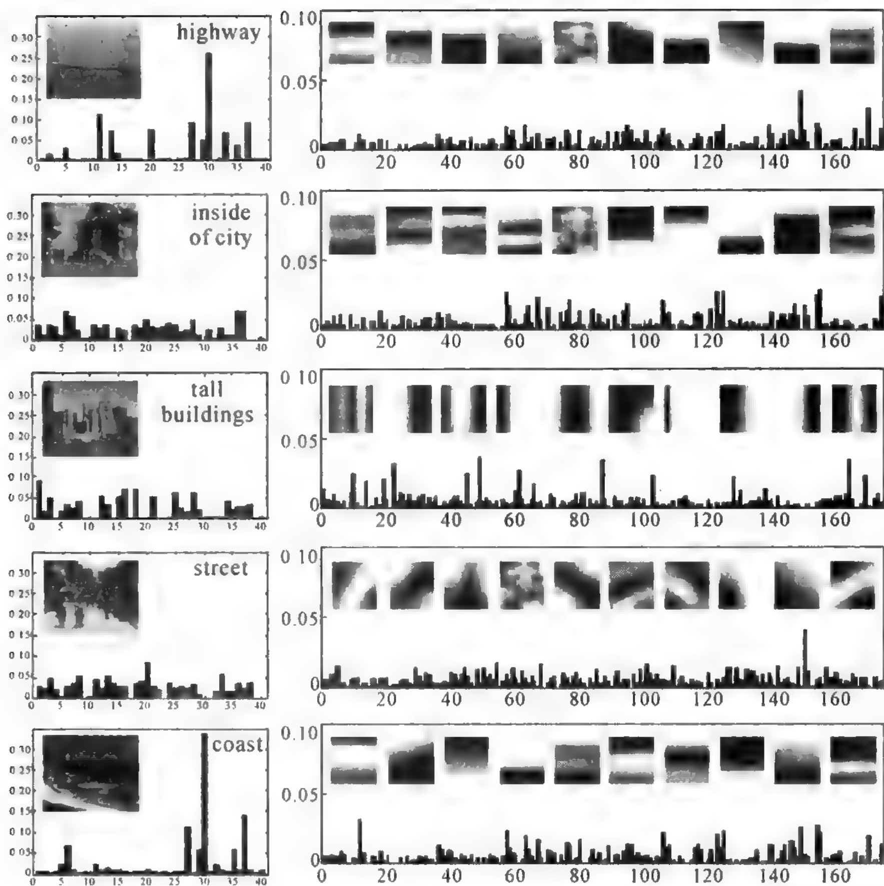


图 1.2　图像分类字典生成效果

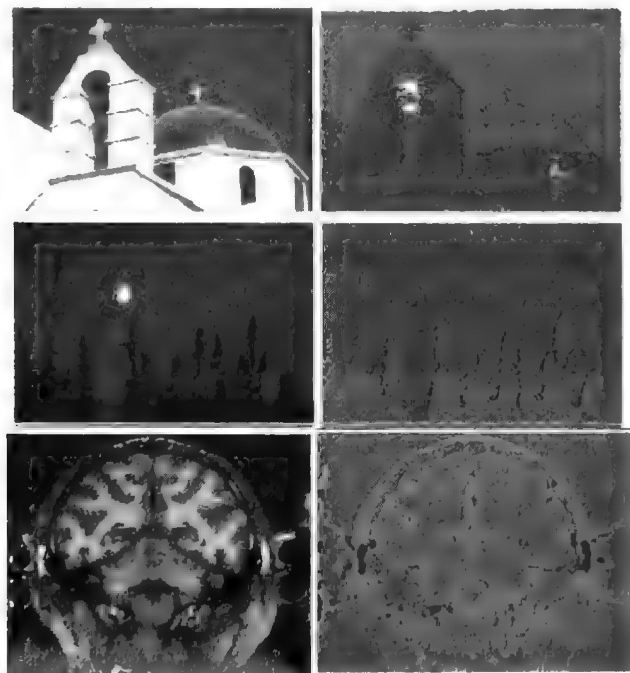


图 1.3　图像分割效果



图 1.4 图像插值效果

综上所述,基于贝叶斯非参数的统计稀疏学习方法是人工智能、应用统计学及视觉认知等学科交叉的研究方向,也是当前统计学习领域的最新研究热点之一,已经引起了国内外学者的重点关注和研究兴趣。其研究成果不仅对统计学习的理论研究具有重要的促进作用,而且在大规模数据挖掘、多媒体内容语义分析、视觉行为自动注解、机器人交互等各种应用领域具有巨大的技术潜力。

1.2 国内外研究现状

1.2.1 统计稀疏学习方法

统计稀疏学习方法的研究起源于多个研究领域的成果:①来源于神经生物学对人类视觉皮层的认知机理研究,这为机器学习提供了生物学上的认知模型借鉴。②来源于数学等领域的最新研究进展,主要包括美国 Stanford 大学统计系的 Donoho 和 Candès、美国 UCLA 数学系的陶哲轩及美国 RICE 大学的 Baraniuk 等人在压缩感知理论和稀疏信号编码方面的开创性工作,为稀疏表达提供了基本的理论依据。③来源于凸优化理论方面的研究进展,主要包括 Stanford 大学的 Michael Saunders 等开发的凸优化算法,为稀疏表达的约束优化求解提供了理论支持和可计算方法。④来源于机器学习领域的理论研究进展,主要包括美国 UC Berkeley 大学统计系的 M. I. Jordan 研究组、Stanford 大学统计系的 Trevor Hastie 研究组等,近年来为稀疏表达与统计学习方法的结合提出了很多有效的学习方法与计算模型。⑤计算机视觉领域的研究者,包括法国 INRIA 的 Jean Ponce、美国 Stanford 大学的 F. F Li、UIUC 的 MaYi 等为稀疏表达在计算机视觉任务上的应用做了很多代表性的工作。

统计稀疏学习研究主要包括三个方面的内容:①稀疏建模,即研究如何构造稀

疏形式，建立有效的稀疏模型。②稀疏求解，即针对特定的稀疏模型，研究如何求解。③稀疏应用，稀疏形式的构造主要通过在最小平均误差项的基础上增加 l_0 范数约束来建模；而稀疏模型的求解则通过约束凸优化方法实现；稀疏应用方面的研究主要针对特定的稀疏问题，研究相应的稀疏学习策略。当前主要的稀疏学习方法基本都涉及上述三个方面的研究，代表性问题包括稀疏降维、字典学习、矩阵填充、稀疏高斯图模型、在线稀疏学习、结构稀疏性等。这些稀疏学习问题虽然已经得到了研究者的关注，但相应的理论和方法还有待研究和完善。

1）稀疏建模及求解

稀疏建模的基本思想是正则化方法，即通过构造同时包括“损失项”和“惩罚项”的约束优化函数来实现。损失项通常采用最小均方误差函数，而惩罚项通常采用范数约束。目前常用的稀疏构造形式包括：l_0 范数约束、Lasso 方法和弹性网（Elastic Net）方法等。从理论上说，l_0 范数约束作为惩罚项具有最优的稀疏形式，但 l_0 范数约束对应 NP 难题，通常无法直接求解。Lasso 方法采用 l_1 范数约束代替 l_0 范数约束构造稀疏模型，由于 Lasso 方法用回归模型系数的绝对值函数作为惩罚来压缩模型系数，使得绝对值较小的系数自动为零，从而实现模型参数选择的自然稀疏性。弹性网方法同时采用 l_1 范数约束和 l_2 范数约束，实现对 Lasso 方法的凸松弛，从而得到较“温和”的稀疏模型；当弹性网方法中 l_2 范数惩罚项的系数为零时，其退化为 Lasso 方法。在一些特定稀疏建模中，不仅对模型系数有稀疏性要求，同时还要求为非负。比如图像像素值的生成，可以在 Lasso 方法或者弹性网方法的基础上，增加对模型系数的非负约束，非负约束的模型来源于 Nonnegative Garrote 方法。在上述几种模型的基础上，Yuan 和 Lin 等提出了 Group-Lasso 方法用于处理结构性稀疏建模，考虑结构性稀疏建模的其他工作还包括 Bach，Huang 等提出的方法。

上述几种稀疏模型构成了典型的凸优化问题，因此能够采用相应的凸优化算法求解。需要注意的是，对于 Lasso 方法，由于 l_1 范数构造的约束是不可微的，这为凸优化问题的求解带来了困难。针对此问题，很多研究者提出了有效的求解策略，典型方法包括内点法（Interior Point Algorithms）、最小角度回归算法（Least Angle Regression，LARS）、正交匹配追踪（Orthogonal Matching Pursuit，OMP）、坐标梯度下降（Coordinate Gradient Descent，CGD）、块坐标梯度下降（Block-Coordinate Gradient Descent，BCGD）等。

2）稀疏降维方法

稀疏降维作为一种有效稀疏编码策略，能够对过完备字典（Overcomplete Dictionary）进行稀疏学习并构造精简的压缩表达。用于视觉任务的稀疏降维，其基函数的选择并不需要采用传统的正交基（傅里叶基、小波基等），而是可直接选自

图像、视频样本中的原始信息或特征表达，这在视觉识别、图像分类等应用中能够有效构造基于内容或语义的信息表达。目前提出的稀疏降维方法是通过对传统降维方法——主元分析法(PCA)增加稀疏性约束实现的，主元分析和主坐标分析(PCO)是统计学习方法中两个重要的无监督降维技术，它们互为对偶问题。因为主元分析降维后的主元包含了所有原始变量的线性组合，难以推断主元与原始变量之间的关系，如果主元只与很少的原始变量相关，则在实际应用中可以为主元与原始变量间建立更易解释的关系。稀疏主元分析(sPCA)通过增加对负荷(loadings)的稀疏约束，比如采用 Lasso 方法或者弹性网方法，构造负荷的自然稀疏性。目前的稀疏降维方法包括两类，一类基于最大主元的协方差性质，比如 Jolliffe 等提出的 SCoTLASS，Shen 等提出的 sPCA-rSVD 等；另一类则是 Zou 等提出的基于回归问题的 sPCA 方法。Zass 等在考虑稀疏性的同时，结合了非负约束，提出了非负主元分析 nsPCA，Jenatton 在考虑变量结构性的条件下提出了结构稀疏主元分析(ssPCA)。

3)稀疏矩阵补充

采用稀疏约束的矩阵补充问题是近年来的研究热点问题，此问题虽然来源于推荐系统研究，但在机器学习和视觉应用中仍然具有广泛的应用价值。矩阵补充通常假设矩阵低秩或近似低秩，并在只有少量观察的情况下，要求恢复矩阵的原始信息，当前有代表性的工作包括 Candès 等提出的近似最优矩阵补充方法和精确矩阵补充方法、Cai 等提出的 SVT(Singular value thresholding)算法，以及 Raghunandan 等提出的从很少观察项中进行矩阵补充的方法。低秩矩阵补充问题主要通过建立最小化目标矩阵的秩并求解相应的约束优化问题。Mazumder 提出采用谱正则化算法(Spectral regularization algorithms)求解矩阵补充问题。Cai 和 Candès 等提出采用核范数(Nuclear norm)惩罚项作为秩约束的凸松弛条件求解。但上述两种方法在收敛速度和恢复精度上都无法同时保持高效。目前，矩阵补充问题的研究还在发展，尤其是对于大矩阵求解时的计算效率，以及当观察矩阵附带噪声时的求解等问题都尚待研究。

统计稀疏学习方法近年来在稀疏约束的矩阵分解、在线字典学习、结构稀疏性等方面有一些新的研究思路，尤其是在统计稀疏学习方法的应用方面，很多研究者做了许多探索性的尝试，并取得初步的研究成果，代表性的应用包括人脸识别、图像超分辨率、图像降噪、背景建模、运动和数据分割及图像分类等。2009 年 Francis Bach 和计算机视觉领域的著名学者 Jean Ponce 等合作，在 ICCV 国际会议上对稀疏编码和字典学习用于图像分析做了系统介绍，MaYi 等从计算机视觉和模式识别的角度也对当前的稀疏表达及其应用做了简介。

目前国内对统计稀疏学习方法的研究也日渐广泛，国内研究者在视觉认知的

编码机制、压缩感知、稀疏表示、字典学习、矩阵分解等的相关理论与应用方面取得了很多重要的研究成果。复旦大学的俞洪波教授等从神经生理学角度对视觉通路的信息编码机制进行研究,南京大学的周志华教授、浙江大学的张志华教授等从统计学习的角度对稀疏学习方法进行了研究。在应用研究方面,杨谦建立了一个基于超定完备基的简单细胞集群稀疏表示的计算模型,实现了自然图像的稀疏编码。中国科学院计算技术研究所李清勇博士设计了面向知觉任务的稀疏编码模型,并扩展了单层的基于 ICA 算法的稀疏编码模型。更多的研究集中于以脊波、曲线波为主线的理论分析及应用、匹配追踪算法在图像处理方面的具体应用、稀疏编码算法在图像处理和图像识别中的应用等。

1.2.2 贝叶斯非参数方法

贝叶斯非参数模型为非参数模型选择和自适应提供了一个贝叶斯框架。然而贝叶斯和非参数方法的结合充满了挑战,因为贝叶斯模型需要明确假设一个在给定参数空间上的概率分布,而非参数模型则根据样本数据改变参数空间的维度。1973 年,Ferguson 在可数样本空间上近似贝叶斯估计的基础上,提出了狄利克雷过程。一方面,既然非参数模型需要一个不受参数数量限制的先验,那么它可以被看作是带有无限维度参数空间的参数模型;另一方面,贝叶斯模型可以通过参数分布来定义,这个参数分布有无限维度的参数空间。这种模型通常称为“贝叶斯非参数模型”。

目前,最常见的贝叶斯非参数模型有高斯过程(Gaussian process)模型和狄利克雷过程(Dirichlet process)模型。高斯过程是传统的多变量高斯分布由向量到函数的自然扩展,其精练的协方差函数结构能极大地降低函数数据分析中的参数估计任务。早在 20 世纪七八十年代,高斯过程就已经以 Kriging 的名义应用于地理统计学领域中,但直到 90 年代中期,经过 Neal, Gibbs 和 MacKay 等人对高斯过程的阐述和发展,高斯过程才受到人们的重视,开始研究应用于机器学习领域,并在各应用领域迅速成为研究的热点。

狄利克雷过程的理论研究在 20 世纪 70 年代是众多研究者关注的热点,研究者们对狄利克雷过程的构造方法、狄利克雷过程的性质、后验计算方法展开了大量的理论研究。但由于其需要大规模的迭代计算,狄利克雷过程的应用一直没有突破性进展。2003 年以来,高速计算机的快速发展解决了边缘概率积分的复杂计算问题,同时,得益于 MCMC 方法、EM 算法、变分方法等的研究,狄利克雷过程的应用研究迅速发展,成为当前机器学习领域的热点。

对贝叶斯非参数方法的研究主要集中在两个方面:模型研究和应用研究。贝叶斯非参数方法以无限维度空间中的随机过程为研究对象,其理论研究包括建模

的方法、模型的性质、模型的推导和演绎等;而应用方面的研究主要针对特定的问题,研究相适应的贝叶斯非参数学习策略。

1)模型研究

贝叶斯非参数方法需要对无限维度空间中的测度进行建模,但在有限维空间中与 Lebesgue 测度相关的密度,如高斯概率密度、狄利克雷分布等,不能直接扩展到无限维空间,需要寻找适合的方法构建贝叶斯非参数模型。Ferguson 通过观察发现,既然非参数模型需要一个不受参数数量限制的先验,那么可以把它看作带有无限维度参数空间的参数模型,贝叶斯模型通过这种参数分布来定义,就得到贝叶斯非参数模型。

对于贝叶斯非参数模型的构造,研究者提出多种方法,主要有基于随机过程的方法、基于 De Finetti 理论的方法、基于 Kolmogorov 扩展理论的方法等。这些方法并不是相互之间完全排他的,例如狄利克雷过程可以通过多种方法构造。

随机过程方法通常适用于生成实数线上或实数区间上的随机概率分布,通过随机过程的非负增量路径来采样,从而描绘累积分布函数。其中,最典型的例子是 Lévy 过程,它是一个递增的过程,其包含的随机变量在一段时间上的概率只与时间段的长度有关。Lévy 过程在贝叶斯非参数中有广泛的应用,其后验在标准化之后得到 Gamma 过程。另外,基于随机过程的贝叶斯非参数模型还有 Griffin 的随机微分方程定义的过程、Küchler 以时间为参数定义具有指数形式的似然函数的随机过程等。

De Finetti 理论阐述了在给定一系列参数的条件下,变量之间是条件独立的。而对于连续的采样空间,这些参数是无限维的。Hewitt 和 Savage 证明了对于一个给定的随机变量序列,它们的混合分布是唯一的。定义一个无限的可交换的随机变量序列,可以通过指定一个生成算法来保证可交换性。例如,Blackwell 和 MacQueen 通过生成模型的混合分布来构建狄利克雷过程,称为无限 Pólya Urn 机制。

Kolmogorov 扩张定理直接从有限维度空间的边缘分布构建无限维度空间中的测度,是 Ferguson 构建狄利克雷过程的理论基础。Ferguson 证明了狄利克雷过程先验能够满足非参数贝叶斯分析的两个基本要求:①在适当的拓扑下,先验分布的支撑要足够大;②给定样本后,后验分布要便于计算。狄利克雷过程的支撑是可测空间上的所有离散概率分布组成的集合,同时,它的后验分布是容易计算的,可表示成先验与经验分布的混合。

狄利克雷过程的性质是狄利克雷分布的性质向无限维空间扩展的结果。Ferguson 在提出狄利克雷过程的同时对狄利克雷过程的诸多性质进行了证明,这奠定了狄利克雷过程的重要理论基础。Orbanz 着重从边缘分布的角度对狄利克

雷过程的性质进行阐述，强调了狄利克雷过程的合并过程。2010年，Subhashis对狄利克雷过程的性质进行了比较全面的总结。在阐述和证明狄利克雷过程性质过程中，Dubins和Pitman根据狄利克雷过程和多项式过程的共轭性，提出“中国餐馆过程(Chinese Restaurant Process)”，在不限定类别数量的前提下对数据的聚类特性进行描述。

2)贝叶斯非参数模型的应用研究

近年来，贝叶斯非参数模型被应用于多种问题，例如回归、分类、聚类、隐变量模型，序列模型，图像分割，信号分离和语法归纳等，其中LDA(Laten dirichlet allocation)模型是最典型也是最成功的贝叶斯非参数模型的应用。其在自然语言和智能信息处理中充分发挥了贝叶斯非参数在无限维度空间建模的优点。模型将主题混合权重视为多维参数的潜在随机变量，推理上采用Laplace近似、变分近似、MCMC(Markov chain Monte Carlo)及“期望-扩散(expectation propagation)”等方法获取待估参数值，在自然语言的词性标注、主题分解、信息抽取等方面取得广泛应用。狄利克雷过程在生物信息处理中也获得了令人惊叹的效果，例如在DNA排序技术中对单倍体型分期(haplotype phasing)用层次狄利克雷过程建模，从而提高了长片段测序能力。对语音的识别是狄利克雷过程的另一应用领域，以“层次狄利克雷过程-隐马尔可夫模型”建模的“说话人检索(Speaker Diarization)”可以在复杂的环境中快速识别说话者。

Sudderth，Li和Paisley等对狄利克雷过程在计算机视觉中的应用进行了一定的探索。Sudderth采用Piteman-Yor过程实现了对图像的分割和标注，并利用高斯过程对Piteman-Yor过程的空间独立性进行描述。Li利用LDA对自然图像进行注解，Paisley等首次利用“贝努利-贝塔过程”描述字典学习并将其应用在图像降噪、图像插值等方面，尽管效果差强人意，但为图像处理提供了新的方法和解决思路。

国内对于贝叶斯非参数方法的理论研究主要集中在20世纪90年代中期，主要对狄利克雷过程的性质、Lévy表示、右中立过程等进行了分析和阐述。然而对贝叶斯非参数模型的应用研究尚处于起步阶段，目前鲜有对贝叶斯非参数模型在应用研究中的综述文献，对贝叶斯非参数中狄利克雷过程、贝塔过程等构造方法的相关研究和应用也亟待发展。目前国内相关研究的论文，有卿湘运等结合狄利克雷过程混合模型和选择特征子集的非参数模型，设计了基于马尔可夫链蒙特卡罗的参数后验推断算法，并将其应用于人脸聚类问题。

第 2 章
贝叶斯非参数模型的构建

自从 Ferguson 在 1973 年提出以带有无限维度参数空间的参数模型来表示先验的方法后，涌现了大量的构建贝叶斯非参数模型的方法。正是基于这些不同的模型构建方法，贝叶斯非参数过程得以广泛地应用在聚类、回归、变量选择等问题中。本章首先介绍贝叶斯非参数的理论基础，在此基础上，分析比较贝叶斯非参数模型的几种构建方法，再针对稀疏表示，对具有稀疏特质的贝叶斯非参数过程的构建和推理方法进行演绎，为全文的研究提供基本的方法。

2.1 符号约定

本书对有关概率模型和随机变量的符号约定如下：

随机变量定义在一个普通、抽象的概率空间 $(\Lambda,\mathcal{A},P)$ 中，其中 Λ 是一个非空集合，有时称为样本空间，$\mathcal{A}$ 是σ-代数，P 是概率或概率测度。随机变量可以从这个普通的概率空间映射到相应的采样空间。随机变量用 X 表示，采样空间和他们的 σ-代数用相应的随机变量作为索引标注。例如，随机变量 $X:(\Lambda,\mathcal{A})\rightarrow(\Omega_x,\mathcal{A}_x)$，其中 Ω_x 是随机变量 X 的采样空间，X 在采样空间中的采样值用小写斜体字母 $\boldsymbol{x}$ 表示。

对于有特定用途的变量，约定用 X 表示观测变量，Θ 表示参数变量，Y 表示超参数。任意的σ- 代数用A，C 等表示，但$\mathcal{B}$表示Borel σ- 代数。随机变量 X 的概率测度 μ 为$\mu = X(P)$。对于同一上下文中的多个随机变量，测度用随机变量作为索引标注，例如 μ_X，μ_Θ。

条件概率记为 $\mu(X\mid\Theta)$，在运算过程中，根据上下文，X 可能被测度集合代替，Θ 可能被σ- 代数代替。概率空间 Λ 中的元素用ω 表示，则把条件概率看作函数的话，可表示为 $\mu(X\mid\Theta)(\omega)$。如果 $\mu(X\mid\Theta)$ 有条件密度，则记为 $p(x\mid\theta)$。字母 s

通常表示充分统计量，$S:=s(X)$ 是随机变量。期望用 E 表示，条件期望表示为 $\mathrm{E}[X \mid C]$。期望也可以用随机变量或测度进行索引标注，例如 $\mathrm{E}_X[\cdot]$，$\mathrm{E}_{\mu(X\mid\Theta)}[\cdot]$。

为避免混淆，一般地，小写斜体表示标量，如 ω_i，t_i 等；小写粗体表示向量，如 $\boldsymbol{x}$，$\boldsymbol{w}$ 等；而大写粗斜体或大写希腊字母表示矩阵，如 $\boldsymbol{A}$，$\boldsymbol{\Phi}$，$\boldsymbol{\Sigma}$ 等。此外，大写 $P(\cdot)$ 表示离散的概率分布函数，而小写 $p(\cdot)$ 则是连续的概率分布函数。

2.2 贝叶斯非参数模型

一个典型的统计问题可以描述如下：首先进行一系列的随机试验，收集样本数据，再对样本数据进行分析、总结，然后进行推断和预测，为相关决策提供依据和参考。在这个描述中，收集样本数据，并对数据进行分析和总结属于描述统计研究的范畴，而推断统计是研究如何根据样本数据去推断总体数量特征的方法，它是在对样本数据进行描述的基础上，对统计总体的未知数量特征做出以概率形式表述的推断。

在推断过程中，如果对总体分布假设的概率模型可以用一系列参数表示，且模型的参数数量不依赖于观测数据的数量，则这个模型是参数模型。非参数模型用一种特别的方法来选择模型和调整自适应性，模型的尺寸随着采样的尺寸增大而增加。例如，参数方法进行密度估计意味着通过最大似然选择一个高斯或者固定数目的混合高斯。而非参数方法是用一个 Parzen Window 估计器，它对于每个观测值集中于一个高斯，因此每个观测值有一个均值参数。

参数模型和非参数模型最基本的不同之处是参数模型倾向于有更加理论化的保证和更快的收敛速度，而非参数模型更适应于需要模型自适应性强的问题。非参数方法在经典统计（非贝叶斯）中已经流行已久，尽管非参数模型的理论结果与参数模型相比很难证明，但它们在应用中的效果让人印象深刻。

贝叶斯模型把参数看作随机变量，模型为每个参数假设一个概率分布，这些概率分布由样本数据来确定。贝叶斯和非参数方法结合充满了挑战，因为贝叶斯模型需要明确假设一个在给定参数空间上的概率分布，而非参数模型根据样本数据改变参数空间的维度。Ferguson 通过观察到如下结果并解决了这个问题：既然非参数模型需要一个不受参数数量限制的先验，那么它可以被看作带有无限维度参数空间的参数模型；贝叶斯模型可以通过参数分布来定义，这个参数分布有无限维度的参数空间。这种模型现在通常称为贝叶斯非参数模型。

定义 2.1：（贝叶斯非参数模型）一个采样模型为 $\mu_X(X \mid \Theta)$、先验分布为 $\mu_\Theta(\Theta)$ 的贝叶斯模型如果满足：存在一个数 $n_0 \in \mathbf{N}$，使得每个额外的观测最多需要

参数空间中 n_0 个额外的参数，且任意观测序列 $\{x_1,\cdots,x_n\}$ 的参数数量的期望值是随着 n 的增加而单调递增的，那么这个贝叶斯模型是贝叶斯非参数模型。

这个模型有足够数量的参数来解释任意一个采样，这也是定义无限维模型的基本原因。如果一个非参数贝叶斯模型有无限个参数，那么它就可以解释任意给定尺寸的采样。为采样尺寸设置一个"无限"的限制是为了在不修改模型的前提下研究模型的渐近行为。几乎所有的贝叶斯非参数模型都是无限维的模型。

贝叶斯非参数模型为非参数模型选择和自适应提供了一个贝叶斯框架。它在一个无限维度的参数空间中，只调用参数的一个有限子集，这个子集通常随着数据集的增加而扩大。在贝叶斯非参数模型的上下文中，"无限维度"可以翻译为"有限的但无界的维度"。贝叶斯非参数模型的关键特征是能够解释局部观测，一次采样涉及的参数只是模型参数的一个子集。

定义 2.2：（*局部观测*）X 是一个有多参数的随机变量，它的值在如下乘积结构的空间中

$$\Omega^E=\prod_{i\in E}\Omega^i \tag{2.1}$$

其中，Ω^i 是任意的部分空间。对于任意 $I\in E$，I 中元素的局部乘积表示为 Ω^I，则受限变量 $X^I=X|_{\Omega^I}$ 的一个观测值。

在贝叶斯非参数模型中，局部观测 X^I 通常表示为有限的观测集合 $\{x_1,\cdots,x_n\}$，I 的尺寸随着 n 的增加而增大。

贝叶斯非参数模型和贝叶斯参数模型之间的区别在于估计的过程如何影响模型的维度。贝叶斯参数模型丢弃了没有被采样数据提及的维度，而贝叶斯非参数模型保留所有的维度，并在先验假设中假设那些不能被观测数据估计的参数。例如，一个采样尺寸为 n 的 Parzen 估计有 n 个精确的本地参数，如果贝叶斯非参数模型的先验假设参数维度为 d，d 可能是有限的也可能是无限的，那么不管采样尺寸如何，一定会估计一个 d 维的后验。

根据定义 2.1 和定义 2.2，贝叶斯非参数模型可以描述为具有如下两个特征的模型：

(1)在一个无限维度的参数空间上构造一个贝叶斯模型；

(2)可以通过有限采样进行求解，求解的方法是只用所有可能解的一个有限子集来解释这些采样。

近年来，贝叶斯非参数模型以其灵活性获得广泛的关注，这种关注尤其表现在非监督学习中。贝叶斯非参数模型的灵活性一方面是模型的表现（参数的个数、参数的结构）能够随着观测数据的增加而增加，另一方面，模型中先验和后验的分布不是参数分布，而是随机过程。贝叶斯非参数模型被应用于多种问题，例如回归、

分类、聚类、隐变量模型、序列模型、信号分离和语法归纳等。

2.3 相关理论基础

贝叶斯非参数模型需要无限维度空间中的测度，那么首要的问题是这样的测度是否存在，如果存在，有多少这样的测度。有限维度空间中的高斯可以写成一个封闭的形式，即与 Lebesgue 测度相关的密度。但因为 Lebesgue 测度不能扩展到无限维空间，所以高斯不能直接扩展到无限维空间。其他的无限维模型也同样不能给出密度表示(包括狄利克雷过程)，需要寻找其他的表示方法构建贝叶斯非参数模型。

自从狄利克雷过程在1973年被提出以来，多种不同原理的贝叶斯非参数模型的构造方法被提出，例如基于随机过程、De Finetti 理论、Kolmogorov 扩展理论等。但这些理论之间并不是完全互斥的，一个非参数过程可以通过多种理论演绎构造方法，例如狄利克雷过程。这些理论是分析和研究贝叶斯非参数模型构造方法的基础，本节对其进行简单介绍，相关的定理在任何测度论的书籍中都可以找到，因此本节不对其进行证明。

2.3.1 随机过程方法

在概率论中，与确定性过程随时间演变只有一个可能的路径不同，随机过程中存在一些未来演变的不确定性，这种不确定性由概率分布来描述。即使初始条件是已知的，随机过程也有多种演变的路径，只是有些路径的概率更高，有些路径的概率更低。

随机过程方法通常适用于生成实数线上或实数区间上的随机概率分布，通过随机过程的非负增量路径来采样，从而描绘累积分布函数(CDF)。例如，Ferguson 给出在区间 $[a,b]$ 上的狄利克雷过程的定义，通过如下步骤生成 CDF：

(1)通过在区间 $[a,b]$ 上的 Gamma 过程采样路径来生成随机函数 f；

(2)对 f 进行标准化 $\bar{f}(x) := \frac{f(x)}{f(b)}$。

通常，CDF 得到的结果往往与实际结果有很大偏差，人们更希望得到概率密度函数，但 CDF 是递增的，能够反映区间的局部特征。例如，Lévy 过程是最典型的独立的递增过程，其在贝叶斯非参数中有广泛的应用。Lévy 过程得到的后验在标准化之后得到 Gamma 过程。

2.3.2 De Finetti 定理

可交换性是贝叶斯非参数模型重要的理论基础。同一概率空间 (Λ, A) 中有 N 个随机变量 $X_1, X_2, \cdots, X_N$，如果这些随机变量的联合分布与变量在序列中的位置无关，则这个变量序列是可交换的，即

$$p(X_1, \cdots, X_N) = p(X_{\tau(1)}, \cdots, X_{\tau(N)}) \tag{2.2}$$

其中 $\tau(\cdot)$ 表示对索引号的任意排列。

当 $N \to \infty$ 时，对于变量序列 $X_1, X_2, \cdots$，如果对于任意 $N \geqslant 1$，$X_1, \cdots, X_N$ 是可交换的，则这个无穷的变量序列也是可交换的。即无穷变量序列的任意有限子集都是可交换的，则该无穷变量序列是可交换的。可交换性表明了随机变量的联合分布不依赖于随机变量之间的位序，但变量之间可能存在依赖性。独立同分布变量是可交换的，但可交换变量不一定是独立同分布的。

定理 2.1：(De Finetti 定理)对于任意无穷可交换的变量序列 $\{X_i\}_{i=1}^{\infty}, X_i \in \Lambda$，存在参数空间 Θ 和相应的分布 $p(\theta)$，使得任意 N 个随机变量的联合分布有如下的混合表示：

$$P(X_1, X_2 \cdots, X_N) = \int_{\Theta} P(\theta) \prod_{i=1}^{N} P(X_i \mid \theta) \mathrm{d}\theta \tag{2.3}$$

当 Λ 是一个 K 维空间时，Θ 是 $K-1$ 单纯形。当 Λ 是欧几里得空间时，Θ 是概率测度的一个无穷维度空间。

可交换性并不意味着随机变量之间的独立性，但根据 De Finetti 理论，在给定一系列参数 θ 的条件下，变量之间是条件独立的。值得注意的是，对于连续的采样空间 Θ，这些参数是无限维的。Hewitt 和 Savage 已经证明，给定一个随机变量序列，它们的混合分布是唯一的。定义一个无限的可交换的随机变量序列，可以通过指定一个生成算法来保证可交换性。例如，Blackwell 和 MacQueen 通过生成模型的混合分布来构建狄利克雷过程，称为无限 Pólya Urn 机制。

2.3.3 Kolmogorov 扩张定理

Kolmogorov 扩张定理直接从有限维度空间的边缘分布构建无限维度空间中的测度。以高斯过程为例，Kolmogorov 扩张定理保证，对于一个随机变量的集合，其任意有限子集的联合分布都是高斯分布，高斯过程测度对于整个集合都是存在的而且是唯一的。类似的，Ferguson 依赖 Kolmogorov 扩张定理，通过有限子集的狄利克雷边缘分布定义了无限维度的测度。

贝叶斯非参数模型是对包含了无限元素的随机对象的概率分布，如何定义这个分布是首先需要考虑的问题。对于随机变量 $X_i, i \in E$，E 是一个无限的索引集

合，那么 X_E 是随机变量构成无限集合。这个随机变量的无限集合可以看成是无限向量、函数、运算、无限图等。对于 E 的唯一要求是不为空，$E^* = \{I \subset E, |I| < \infty\}$ 表示 E 的所有有限子集。随机变量 X_t 在 Polish 空间 $\Omega^{(t)} = \Omega$ 取值，无限维度的随机变量的采样空间是无限乘积空间 $\Omega^E = \prod_{t \in E} \Omega$，形象地说，$\Omega^E$ 是同一空间 Ω 的无限重复。无限维度测度的有限维边缘分布是 Ω^E 有限维子空间上的边缘分布。对于索引集合的有限子集 $I \subset E^*$，采样空间为 Ω^I，每个随机变量 X^I 的边缘测度为 μ^I。Ω^E 任意的两个子空间 Ω^I 和 Ω^J，$I \subset J$，$x^J = (x_i)_{i \in J}$ 是子空间 Ω^J 中一系列元素，子空间 Ω^J 在 Ω^I 的投影运算为 $P_{J,I} x^J = (x_i)_{i \in I}$，即投影运算是从子空间 Ω^J 中移除那些不在空间 Ω^I 上的元素。$P_{J,I}$ 的预映射用 $R_{J,I}$ 表示，$R_{J,I} x^I = \{x^J \in \Omega^J \mid P_{J,I} x^J = x^I\}$。

定义 2.3：（投影族，Projective Family）令 $\{\mu^I \mid I \in E^*\}$ 是空间 $(\Omega^I, \mathcal{B}^I)$ 上的概率测度族，如果对于任意 $I, J \in E^*$，$I \subset J$，如果满足

$$P_{J,I} \mu^J = \mu^I \tag{2.4}$$

则称此概率测度族为投影族。

假设已经给定无限随机变量 X^E 的测度 μ^E，如果它所有的边缘分布都在 Ω^E 的有限维子集上计算，这些边缘分布在如下意义上一致：如果有 $J, K \in E^*$ 是两个部分重叠的索引集合，令 I 是一个普通子集，$I \subset J$ 并且 $I \subset K$，则 μ^J 和 μ^K 在 Ω^I 上的边缘分布是确定的。如果把边缘分布作为投影，则边缘分布的关系就是式(2.4)。换一种说法，定义 2.3 表明投影族是一个测度系统，它能够构成普通测度 μ^E 的边缘分布。如果这样的测度存在，Kolmogorov 定理表明向下的投影是可逆的，即如果这些测度是投影的，则测度 μ^E 存在并且是唯一的。

定理 2.2：（Kolmogorov 扩张定理）令 $\{\mu^I \mid I \in E^*\}$ 是空间 $\Omega^I, \mathcal{B}^I$ 上概率测度的投影族，则在空间 $(\Omega^I, \mathcal{B}^I)$ 上存在唯一测度 μ^E，它的边缘分布是测度 μ^I。

由 Kolmogorov 扩张定理定义的测度称为投影族 $\{\mu^I \mid I \in E^*\}$ 的投影极限(projective limit)。简单地说，Kolmogorov 扩张定理描述了对于无限维随机变量序列 X，如果它的所有有限子集的边缘分布是已知的，那么它的联合分布就已经被完全定义了。

随机过程、De Finetti 定理和 Kolmogorov 扩张定理是本文的理论基础，在后续章节中，贝叶斯非参数模型构造方法的分析和研究均在这三种理论基础之上展开。

2.4 狄利克雷过程

狄利克雷过程是狄利克雷分布在无限维度中的扩展，本节从有限维度的狄利

克雷分布开始，到无限维度的狄利克雷过程。

2.4.1 狄利克雷分布

当 $\chi=\{1,\cdots,K\}$，从 K 个离散的类中取一个随机变量 X，如果第 k 类占总体的比例为 π_k，则 X 的概率为：

$$p(X \mid \boldsymbol{\pi})=\prod_{k=1}^{K}\pi_k^{\delta(x,k)}$$

其中 $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_K)^T,\pi_k\geqslant 0$，且 $\sum_k \pi_k=1$。当 x 属于第 k 类时，$\delta(x,k)=1$，否则 $\delta(x,k)=0$。向量 $\boldsymbol{\pi}$ 被称为是取自 $\mathbf{R}^K$ 的 $K-1$ 维单纯型空间的向量，表示为 $\boldsymbol{\pi}\in\Delta_K$。如果有 L 个观测值 $\{X^{(l)}\}_{l=1}^{L}$，则这 L 个观测值的联合概率分布为：

$$p(X^{(1)},\cdots,X^{(L)} \mid \pi_1,\cdots,\pi_K)=\frac{L!}{\prod_k C_k!}\prod_{k=1}^{K}\pi_k^{C_k}$$

其中 $C_k=\sum_{l=1}^{L}\delta(X^{(l)},k)$。

当有 L 个观测值时，可以对多项式参数 $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_K)$ 进行最大似然估计：

$$\hat{\boldsymbol{\pi}}=\underset{\boldsymbol{\pi}}{\operatorname{argmax}}\sum_{l=1}^{L}\log p(X^{(l)} \mid \boldsymbol{\pi})=\left(\frac{C_1}{L},\cdots,\frac{C_K}{L}\right) \tag{2.5}$$

但是，当观测数 L 不大于类别数 K 时，根据最大似然估计得到的 $\hat{\boldsymbol{\pi}}$ 中有很多的 0 值，从而导致错误估计。

贝叶斯方法通常假设其为某共轭先验分布，从而进行贝叶斯推理，多项式分布的共轭先验分布是狄利克雷分布。

1)狄利克雷分布的定义

K 维狄利克雷分布是一个连续的概率分布，其密度函数为：

$$p(\boldsymbol{\pi} \mid \boldsymbol{\beta})=\frac{\Gamma\left(\sum_{k=1}^{K}\beta_k\right)}{\prod_k\Gamma(\beta_k)}\prod_{k=1}^{K}\pi_k^{\beta_k-1} \tag{2.6}$$

其中 $\beta_k\geqslant 0,\forall k$，通常记为 $\mathrm{Dir}(\boldsymbol{\pi}\mid\boldsymbol{\beta})$。如果 K 个参数都相等，那么就在 $K-1$ 维的单纯型空间中均匀分布，如果参数不相等，那么就会偏向，例如图 2.1 描述的三维狄利克雷分布，其参数分别为 $\boldsymbol{\beta}=(2,2,2),(6,2,2),(3,7,5),(6,2,6)$。

将狄利克雷分布中的参数 β_k 用 αg_{0k} 表示，其中，$\alpha=\sum_k\beta_k$，$g_{0k}=\frac{\beta_k}{\sum_k\beta_k}$，则狄利克雷分布表示为：

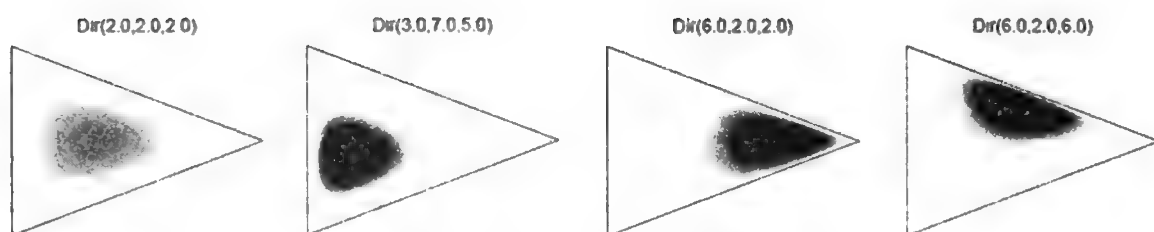


图 2.1　狄利克雷分布

$$p(\boldsymbol{\pi} \mid \alpha g_{01}, \cdots, \alpha g_{0K}) = \frac{\Gamma(\alpha)}{\prod_k \Gamma(\alpha g_{0i})} \prod_{k=1}^{K} \pi_k^{\alpha g_{0k}-1} \tag{2.7}$$

记 $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha \boldsymbol{g}_0)$，$\alpha$ 被称为集中参数(Concentrate Parameter)，$\boldsymbol{g}_0 = (g_{01}, g_{02}, \cdots, g_{0K})$ 称为期望参数(Expectation Parameter)。

通过这种参数变形后，$\boldsymbol{\pi}$ 中各分量的均值和方差为：

$$\mathrm{E}[\pi_k \mid \alpha \boldsymbol{g}_0] = g_{0k}, \ \mathrm{Var}[\pi_k \mid \alpha \boldsymbol{g}_0] = \frac{g_{0k}(1-g_{0k})}{\alpha(\alpha+1)} \tag{2.8}$$

因此，$\boldsymbol{g}_0$ 可以看作 $\boldsymbol{\pi}$ 的先验假设，当 $g_{0k}=0$，则必然有 $\pi_k=0$。当 $g_{0i}=1, g_{0k}=0, \forall k \neq i$，则以概率 1 得到 $\boldsymbol{\pi}=\mathrm{e}_i$($\mathrm{e}_i$ 为单位向量)。α 是集中参数，表示实际分布与期望参数 $\boldsymbol{g}_0$ 的紧密程度。

狄利克雷分布是多项分布的共轭分布函数，通常作为多项分布的共轭先验。多项式分布的概率公式为：

$$P_{Mult}(\boldsymbol{h} \mid \boldsymbol{\pi}) = \frac{(\sum_j h_j)!}{\prod_j h_j!} \exp\left(\sum_{j=1}^{k} h_j \log(\pi_j)\right) \tag{2.9}$$

例如，对于某随机实验可能的结局有 K 种，分别是 $A_1, A_2, \cdots, A_k$，它们出现的概率分布分别是 $\pi_1, \pi_2, \cdots, \pi_K$，则在 N 次采样的总结果中，A_1 出现 h_1 次，A_2 出现 h_2 次，…，A_k 出现 h_k 次，则多项式概率公式(2.9)表示了这样的事件出现的概率。

2)狄利克雷分布的后验计算

狄利克雷分布是多项式分布的共轭先验，根据指数族的性质，$\boldsymbol{\pi}$ 的后验分布也是一个狄利克雷分布，即如果先验为 $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha \boldsymbol{g}_0)$，某次实验观测值 $X \in A_i$，要计算后验 $p(\boldsymbol{\pi} \mid X=i, \alpha \boldsymbol{g}_0)$，根据贝叶斯公式，得到：

$$\begin{aligned} p(\boldsymbol{\pi} \mid X=i, \alpha \boldsymbol{g}_0) &= \frac{p(X=i \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \alpha \boldsymbol{g}_0)}{\int_{\pi \in \triangle_k} p(X=i \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \alpha \boldsymbol{g}_0) \mathrm{d}\pi} \\ &\propto p(X=i \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \alpha \boldsymbol{g}_0) \end{aligned} \tag{2.10}$$

在给出观测值 X 的条件下,可以计算 $\boldsymbol{\pi}$ 的后验分布,首先,观测值 $X=i$ 的概率为 $p(X=i\mid\boldsymbol{\pi})=\pi_i$,将其与先验 $p(\boldsymbol{\pi}\mid\alpha\boldsymbol{g}_0)$ 相乘,则得到后验为:

$$p(\boldsymbol{\pi}\mid X\in A_i,\alpha\boldsymbol{g}_0)\propto\pi_i\prod_{k=1}^{K}\pi_k^{\alpha g_{0k}-1}\tag{2.11}$$

这个表达式与 $\mathrm{Dir}(\alpha\boldsymbol{g}_0+\mathbf{e}_i)$ 成比例,即把先验的狄利克雷分布在第 i 个参数上增加了 1。

以此类推,当观测数据有 L 个时,有

$$\begin{aligned}p(\boldsymbol{\pi}\mid X_1=\boldsymbol{x}_1,\cdots,X_L=\boldsymbol{x}_L)&\propto\prod_{l=1}^{L}p(X_i=\boldsymbol{x}_i\mid\boldsymbol{\pi})p(\boldsymbol{\pi}\mid\alpha\boldsymbol{g}_0)\\&=\prod_{k=1}^{K}\pi_k^{\alpha g_{0k}+C_k-1}\end{aligned}\tag{2.12}$$

其中 C_k 表示观测数据为第 k 个分量的个数且 $\sum_{k=1}^{K}C_k=L$。标准化后,后验分布为

$$\mathrm{Dir}(\alpha g_{01}+C_1,\cdots,\alpha g_{0K}+C_K)\tag{2.13}$$

在给多项式分布一个共轭先验后,$\boldsymbol{\pi}$ 的后验也是一个狄利克雷分布,只是其中的参数通过统计观测数据来获得更新,这种先验和观测数据的互动可以通过后验分布的均值来反映:

$$\mathrm{E}[\pi_k\mid X_1=\boldsymbol{x}_1,\cdots,X_L=\boldsymbol{x}_L]=\frac{\alpha g_{0k}+C_k}{\alpha+N}=\frac{\alpha g_{0k}}{\alpha+N}+\frac{C_k}{\alpha+N}\tag{2.14}$$

式(2.14)直观地反映了先验和观测值对后验的贡献。与式(2.5)相比,式(2.14)表示估计的参数被狄利克雷先验平滑了。

3) *Dirichelt* 分布的生成

有多种方法可以生成符合参数 $\alpha\boldsymbol{g}_0$ 的狄利克雷分布。最常见的方法是由 *Gamma* 分布来构造。

首先生成符合 *Gamma* 分布的变量 Z_i,且

$$Z_i\sim\mathrm{Gamma}(\alpha g_{0i},\lambda)\tag{2.15}$$

其中 αg_{0i} 是形状参数,λ 是尺度参数,可以是任意正常数,$i=1,\cdots,K$。

令

$$\boldsymbol{\pi}:=(\frac{Z_1}{\sum_i Z_i},\cdots,\frac{Z_K}{\sum_i Z_i})\tag{2.16}$$

得到的 $\boldsymbol{\pi}$ 是符合参数为 $\alpha\boldsymbol{g}_0$ 的狄利克雷分布。

可以利用 *Beta* 分布来构造,方法如下:

$$V_k \sim \text{Beta}(\alpha g_{0k}, \alpha \sum_{l=k+1}^{K} g_{0l})$$

$$\pi_k = V_k \prod_{l=1}^{k-1} (1 - V_l) \tag{2.17}$$

$$\pi_K = 1 - \sum_{k=1}^{K-1} V_k$$

得到的 $(\pi_1, \cdots, \pi_K) \sim \text{Dir}(\alpha \boldsymbol{g}_0)$，证明见 *Paisley* 在 2010 年的论文。在无限维度空间中采用这种构造方式则得到狄利克雷过程中的 *Stick-breaking* 过程。

另一种构造狄利克雷分布的方法是通过 *Pólya Urn* 过程来构造，*Pólya Urn* 过程以一个序列的方式获得具有狄利克雷先验的随机离散概率分布的采样，采样过程描述如下：假设有一个罐子，里面有 α 个球，这些球共有 K 种颜色，其中第一种颜色的球有 αg_{01} 个，第二种颜色的球有 αg_{02} 个，以此类推，$\boldsymbol{g}_0 \in \triangle_K$。随机从罐子中取一个球 X_1，这个球的颜色是第 x_1 种的概率是 g_{0x_1}，然后把这个球放回罐子，并在罐子中放一个相同颜色的球。每次取球之后，罐子中都增加一个球，这个过程重复 N 次。

计算这一过程中拿出球的颜色的概率。在第一次取球和放球之后，第二次拿出的球的颜色概率为：

$$p(X_2 \mid X_1 = l) = \frac{1}{\alpha + 1} \delta_{x_1} + \frac{\alpha}{\alpha + 1} \sum_{k=1}^{K} g_{0k} \delta_k \tag{2.18}$$

则在 N 次取球放球之后，取得的第 $N+1$ 个球的颜色的概率为：

$$\begin{aligned} & p(X_{N+1} \mid X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N) \\ &= \sum_{k=1}^{K} \frac{C_k}{\alpha + N} \delta_k + \frac{\alpha}{\alpha + N} \sum_{k=1}^{K} g_{0k} \delta_k \end{aligned} \tag{2.19}$$

第 $N+1$ 个球的颜色为 x_k 的概率为

$$\begin{aligned} & p(X_{N+1} = x_k \mid X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N) \\ &= \frac{C_k}{\alpha + N} + \frac{\alpha}{\alpha + N} g_{0k} \end{aligned} \tag{2.20}$$

根据 $\int_{\Omega_n} p(A \mid B) p(B \mid C) \mathrm{d}B = p(A \mid C)$，有

$$\begin{aligned} & p(X_{N+1} \mid X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N) \\ &= \int p(X_{N+1} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N) \mathrm{d}\boldsymbol{\pi} \end{aligned}$$

在给定 π 的条件下，第 $N+1$ 取得第 k 种颜色球的概率为 $p(X_{N+1} = x_i \mid \boldsymbol{\pi}) = \pi_i$，所以有

$$
\begin{aligned}
&p(X_{N+1} = x_k \mid X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N) \\
&= \int p(X_{N+1} = x_k \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N) \mathrm{d}\boldsymbol{\pi} \\
&= \int \pi_k p(\boldsymbol{\pi} \mid X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N) \mathrm{d}\boldsymbol{\pi} \\
&= \mathrm{E}[\pi_k \mid X_1 = x_1, X_2 = x_2, \cdots, X_N = x_N] \\
&= \frac{C_k}{\alpha + N} + \frac{\alpha g_{0k}}{\alpha + N}
\end{aligned}
$$

与狄利克雷分布的后验(2.14)一致。在无限维度空间中采用 *Pólya Urn* 过程则得到有名的中国餐馆过程。

2.4.2 狄利克雷过程

非参数方法用随机过程对无限维度空间进行建模，常常用在有限维度空间的相应分布进行描述。例如高斯过程的定义，一个函数 $f:\chi \to \mathbf{R}$ 是依据高斯过程的分布，当且仅当任意 N 个点 $x_i \in \chi$ 的函数值的概率密度 $p(f(x_1), \cdots, f(x_N))$ 是联合高斯分布。所以高斯过程可以以均值函数和方差核作为参数。与高斯过程类似，由 *Kolmogorov* 扩张定理，狄利克雷过程是有限维狄利克雷分布的投射族。由狄利克雷分布获得的随机采样是有限概率分布，其值是将采样空间 Ω 分割成有限数量的柱状图(*Histogram Bins*)。有限的索引子集 $I \in E^*$ 表示把 Ω 分为有限个数的子集。对于测度空间 (Ω, A)，将 Ω 分割为测度集合称为测度分割(*measure partition*)。如果对于 $\forall i, A_i \in A, i \neq j, A_i \cap A_j = \varnothing, \bigcup_{i=1}^{n} A_i = \Omega$，则 $H = (A_1, \cdots, A_n)$ 是一个测度分割。所有的 A-测度分割用 $\mathcal{H}$ 表示，$\mathcal{H}^*$ 表示 $\mathcal{H}$ 中的有限集合。

定义 2.4:(狄利克雷过程)包含概率测度 G_0 的测度空间 $(\Omega, \mathcal{A})$，对于任意的测度分割 $H \in \mathcal{H}(\mathcal{A})$，令 $\mathbf{R}^H$ 是积空间 $\mathbf{R}^H = \prod_{A \in H} \mathbf{R}$，$\mathrm{Sim}(\mathbf{R}, H)$ 是单位单纯型。用 $p_{Dir}^{H}(\cdot \mid \alpha, g)$ 表示 $\mathrm{Sim}(\mathbf{R}, H)$ 上的狄利克雷密度，收敛参数 $\alpha \in \mathbf{R}_+$，期望参数 $g \in \mathrm{Sim}(\mathbf{R}, H)$。对每个测度分割 $H \in \mathcal{H}(\mathcal{B})$，定义属于 $\mathrm{Sim}(\mathbf{R}, H)$ 的向量 g^H 为：$\forall A_i \in H$，$g_i^H := G_0(A)$。密度函数 $p_{Dir}^{H}(\cdot \mid \alpha, g)$ 指定的测度为 μ^H，投影族 $\{\mu^H \mid H \in \mathcal{H}(\mathcal{B})\}$ 的投影极限是基础测度为 G_0 的狄利克雷过程，记为 $\mathrm{DP}(\alpha, G_0)$。

简单来说，狄利克雷过程是定义在随机概率测度上的分布，其参数是 Ω 上的一个基本测度 G_0 和一个作为收敛参数的正标量 α。对空间 Ω 的任意有限分割 $(T_1, \cdots, T_K)$，有

$$
(G(T_1), \cdots, G(T_K)) \sim \mathrm{Dir}(\alpha G_0(T_1), \cdots, \alpha G_0(T_K)) \tag{2.21}
$$

则 G 是一个狄利克雷过程。

根据狄利克雷过程的定义(2.21)和狄利克雷分布的均值表示(2.8)，对于任意区域 $T \subset \Omega$，狄利克雷过程的一个随机采样的测度均值为：

$$\mathrm{E}[G(T)] = G_0(T),\ G \sim \mathrm{DP}(\alpha, G_0) \tag{2.22}$$

即基本测度 G_0 指定 $\mathrm{DP}(\alpha, G_0)$ 的均值，参数 α 与狄利克雷分布的精度参数类似，决定采样相对于基本测度的平均偏离。

2.4.3 狄利克雷过程的性质

狄利克雷过程的性质是狄利克雷分布的性质向无限维空间扩展的结果。*Ferguson* 在给出狄利克雷过程定义的基础上，对狄利克雷过程的诸多性质进行了证明，*Subhashis* 在 2010 年对狄利克雷过程的性质重新进行了比较全面的总结。本章根据后文的需要，仅对狄利克雷过程的共轭性、集中性进行说明。

1) 共轭性

在上一节中，狄利克雷分布是多项式分布的共轭先验。如果把这两个分布都扩展到无限维空间，将多项式模型的投影极限与狄利克雷模型的投影极限相对应，从而得到无限维多项式过程，这就是 *Dubins* 和 *Pitman* 提出的“中国餐馆过程”。

类似于有限狄利克雷分布计算的后验(2.11)，对测度空间的任意有限分割 $(T_1, \cdots, T_K)$，当有观测值 $\bar{\pi} \in T_k$，存在后验密度函数：

$$p((G(T_1), \cdots, G(T_K)) \mid \bar{\pi} \in T_k) = \mathrm{Dir}(\alpha G_0(T_1), \cdots, \alpha G_0(T_k) + 1, \cdots \alpha G_0(T_K)) \tag{2.23}$$

当有 N 个相互独立的采样，则有如下定理。

定理 2.3： $G \sim \mathrm{DP}(\alpha, H)$ 是符合狄利克雷过程的随机测度。给定 N 个相互独立的采样 $\bar{\pi}_i \sim G$，后验测度也是一个狄利克雷过程：

$$p(G \mid \bar{\pi}_1, \cdots, \bar{\pi}_N, \alpha, G_0) = \mathrm{DP}\Big(\alpha + N, \frac{1}{\alpha + N}(\alpha G_0 + \sum_{i=1}^{N} \delta_{\bar{\pi}_i})\Big) \tag{2.24}$$

证明可参见 *Ferguson* 对其直接用有限狄利克雷后验分布的共轭形式进行的证明，也可以参看 *Sethuraman* 的另一种更简单的证明。

既然狄利克雷过程的后验仍旧是一个狄利克雷过程，根据式(2.22)，得到后验期望为：

$$\mathrm{E}[G \mid \bar{\pi}_1, \cdots, \bar{\pi}_N, \alpha, G_0] = \frac{1}{\alpha + N}(\alpha G_0 + \sum_{i=1}^{N} \delta_{\bar{\pi}_i}) \tag{2.25}$$

对于任意 $T \subset \Theta$，其后验期望为：

$$\mathrm{E}[G(T) \mid \bar{\pi}_1, \cdots, \bar{\pi}_N, \alpha, G_0] = \frac{1}{N}(\alpha G_0(T) + \sum_{k=1}^{K} N_k \delta_{\pi_k}(T)) \tag{2.26}$$

其中 $N_k \triangleq \sum_{i=1}^{N} \delta(\bar{\pi}_i, \pi_k), k = 1, \cdots, K$。值得注意的是，$K$ 是一个随机变量，不是一

个固定的值。

由定理(2.24)，在已有 i 个相互独立的采样 $\overline{\pi}_1,\cdots,\overline{\pi}_i\sim G$ 的条件下，第 $i+1$ 个采样可由狄利克雷采样公式得到：

$$\overline{\pi}_{i+1}\mid\overline{\pi}_1,\cdots,\overline{\pi}_i\sim\frac{\alpha}{\alpha+i}G_0+\frac{1}{\alpha+i}\sum_{j=1}^{i}\delta_{\overline{\pi}_j}\tag{2.27}$$

这个过程称为 *Pólya Urn* 机制。

2)集中性

狄利克雷分布中集中参数 α 表示实际分布与期望参数 $\boldsymbol{g}_0$ 的紧密程度。图 2.2 显示了从不同的三维狄利克雷分布获得的 10 000 个采样，图中狄利克雷分布的 $\boldsymbol{g}_0$ 为均匀分布，$\alpha=1,3,10$。当 $\alpha=K$，即与狄利克雷分布的维度相同时，采样点均匀地分布在 $K-1$ 单纯型上。当 $\alpha>K$，采样点集中在 $\boldsymbol{g}_0$ 周围，且 α 越大，越向 $\boldsymbol{g}_0$ 集中。当 $\alpha<K$，采样点集中在 Δ_K 的角落，且 α 越小，采样越向单纯型的角落集中。通常，当 α 与 K 的比越小，则生成的 $\boldsymbol{\pi}$ 就会更稀疏，这种情形与狄利克雷过程有关系。

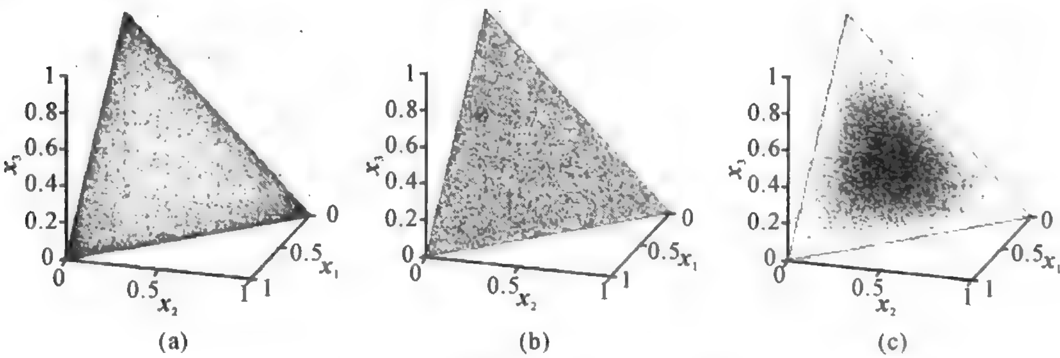


图 2.2　三维狄利克雷分布的投影

在指数族模型中，较大的集中参数使得测量的中心紧密地围绕着它的期望值。在狄利克雷过程中，如果 α 比较大，则由式(2.27)获得的采样多数来自 G_0，从而使得整体经验分布集中于 G_0。但这种数据特征必须在有足够多的采样时才能够反映出来。对于较小的 α，随机测度将集中于最初的几个观测值得到的 *Dirac* 测度，这是因为 *Dirac* 测度表示无限维概率单纯型中的极值点。

表 2.1 显示了狄利克雷过程的集中参数与样本总数之间的关系。样本总数 N 分别为 20,50,100,200,1000，集中参数 α 分别取 N 的 $-2.0\sim3.0$ 整数倍。从表中可以看出，当 α 等于或小于 N^{-1}，得到的聚类个数更倾向于 1；当 α 等于或大于 N^2 时，得到的聚类数更倾向于 N。

表 2.1　基于不同集中参数与样本数得到的聚类数

	N				
	1 000	20	50	100	200
$N^{3.0}$	19.98	49.99	100.00	200.00	1 000.00
$N^{2.0}$	19.54	49.52	99.51	199.50	999.50
$N^{1.0}$	14.12	34.91	69.57	138.88	693.40
$N^{0.0}$	3.60	4.50	5.19	5.88	7.49
$N^{-1.0}$	1.17	1.09	1.05	1.03	1.01
$N^{-2.0}$	1.01	1.00	1.00	1.00	1.00

2.5 狄利克雷过程的构造

对于狄利克雷分布来讲，“构造”是如何得到这 K 个值在 0～1 之间且和为 1 的概率。对于狄利克雷过程来讲，K 是个不确定的数，“构造”是如何得到不确定 K 个值在 0～1 之间且和为 1 的概率。基于 2.3 节中阐述的随机过程、De Finetti 理论和 Kolmogorov 扩展理论等，研究者针对狄利克雷过程设计了多种构造的方法，下面对本书将用到的狄利克雷构建过程进行分析，并给出其与稀疏表示的关系。

2.5.1 Stick-breaking 过程

Sethuraman 提出 Stick-breaking 过程，并证明其是狄利克雷过程。该过程以形象性和易理解性在狄利克雷过程的应用中经常被使用。

Stick-breaking 过程可以看作将一个单位长度的棍子不断折断的过程，且折断次数是无穷的。首先，选择一个位置 $\upsilon_1 \sim \text{Beta}(1,\gamma)$，把折断的部分 υ_1 分配给一个随机的点 $\theta_1 \sim H$。对于剩下的部分 $1-\upsilon_1$，选择一个位置 $\upsilon_2 \sim \text{Beta}(1,\gamma)$ 位置折断，把 $(1-\upsilon_1)\upsilon_2$ 分配给一个随机的点 $\theta_2 \sim H$，再对剩下的部分 $(1-\upsilon_1)(1-\upsilon_2)$ 重复上述的操作。图 2.3 对此过程进行了描述。

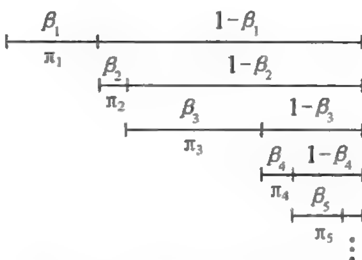


图 2.3 Stick-breaking 过程

这个过程的数学表示如下：

$$
\begin{aligned}
P &= \sum_{i=1}^{\infty}\beta_i\delta_{\theta_i} \\
\theta_i &\overset{\text{i.i.d}}{\sim} H \\
\beta_i &= \upsilon_i\prod_{j=1}^{i-1}(1-\upsilon_j) \\
\upsilon_i &\overset{\text{i.i.d}}{\sim} \text{Beta}(1,\gamma)
\end{aligned} \tag{2.28}
$$

由式(2.28)可得

$$
\begin{aligned}
P &= Y_1\delta_{\theta_1} + (1-Y_1)[Y_2\delta_{\theta_2} + (1-Y_2)Y_3\delta_{\theta_3} + \cdots] \\
&= Y_1\delta_{\theta_1} + (1-Y_1)P
\end{aligned} \tag{2.29}
$$

这种描述形式使得 P 可以通过 MCMC 采样生成。

Stick-breaking 过程具有天然的层次特性，可以表达复杂的层次狄利克雷过程。如果基础测度 G_0 服从狄利克雷过程，以 Stick-breaking 过程对其进行构造，则有

$$G_0 = \sum_{i=1}^{\infty}\beta_i\delta_{\theta_i} \tag{2.30}$$

其中，β_i，θ_i 如式(2.28)描述。

Pitman 将无限序列 $(\beta_1,\beta_2,\cdots)$ 的联合分布称为 GEM(α)。如果随机测度 G_k 也服从狄利克雷过程，而且每个 G_k 的元素都是 G_0 的元素，则 G_k 的 Stick-breaking 表示是：

$$G_k = \sum_{i=1}^{\infty}\pi_{ki}\delta_{\theta_i} \tag{2.31}$$

式(2.31)将问题转变为权重 $\boldsymbol{\beta} = (\beta_1,\beta_2,\cdots)$ 和 $\boldsymbol{\pi}_k = (\pi_{k1},\pi_{k2},\cdots)$ 之间的关系。这些权重向量都是离散空间 $\{1,\cdots,\infty\}$ 的概率测度，以狄利克雷过程对空间的分割来表示对整数的分割，根据狄利克雷过程的收敛性，有：

$$\boldsymbol{\pi}_k \mid \alpha,\boldsymbol{\beta} \sim \text{DP}(\alpha,\boldsymbol{\beta}) \tag{2.32}$$

则对于 π_k 的构造可以表示为：

$$
\begin{gathered}
\upsilon_{ki} \mid \alpha, \beta_1, \cdots, \beta_i \sim \mathrm{Beta}\Big(\alpha\beta_i, \alpha\big(1-\sum_{j=1}^{i}\beta_j\big)\Big),\ i=1,\cdots,\infty \\
\pi_{ki} = \upsilon_{ki}\prod_{j=1}^{i-1}(1-\upsilon_{kj})
\end{gathered}
\tag{2.33}
$$

2.5.2 中国餐馆过程

中国餐馆过程描述了狄利克雷过程的条件分布。如果 $G \sim \mathrm{DP}(\alpha, H)$ 是符合狄利克雷过程的随机测度，在已知 G 生成的 i 个相互独立的采样 $\pi_i \sim G$ 的条件下，第 $i+1$ 个采样的过程称为中国餐馆过程，即

$$
\pi_{i+1} \mid \pi_1, \cdots, \pi_i \sim \mathrm{DP}(\alpha, H) \tag{2.34}
$$

根据狄利克雷采样公式(2.27)，第 $i+1$ 个采样以概率 $\frac{m_k}{\alpha+i}$ 从已有的第 k 个堆中获得，其中 m_k 是已有的 i 个采样中落入该堆的采样的个数；第 $i+1$ 个采样以概率 $\frac{\alpha}{\alpha+i}$ 从新的堆中获得。这个过程相当于第一个顾客进入餐馆后，随机地选择一张桌子坐下；第二个顾客进入后以概率 $\frac{1}{\alpha+1}$ 坐在第一个顾客的桌子旁，以概率 $\frac{\alpha}{\alpha+1}$ 选择新的桌子坐下；随着顾客数量的增加，选择新桌子的概率逐渐降低，顾客更多地集中在人多的桌子旁。图 2.4 显示了已有 4 个顾客时，第 5 个顾客选择各桌子的概率。

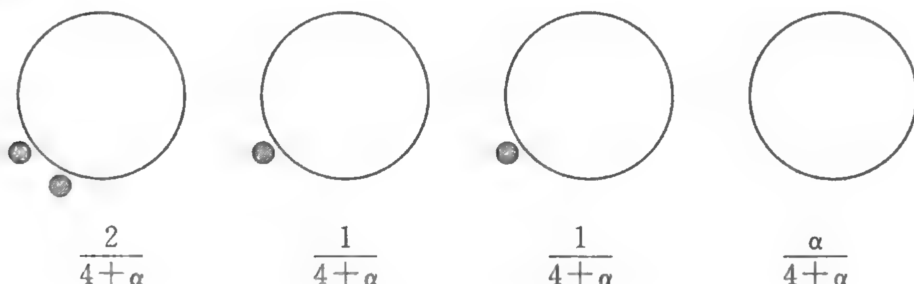


图 2.4　中国餐馆过程

中国餐馆过程描述了聚类的过程，并且能够实现快速的聚类，因为随着顾客总数的增加，更多的顾客聚集在已选的桌子旁，这对于单特征的数据表示非常有利。在本书第 5 章中，正是利用中国餐馆过程的这一特点，对视频背景信息进行建模，从而实现背景剪除。

2.5.3 Pitman-Yor 过程

Pitman-Yor 过程是一个二参数的狄利克雷过程：

$$G \sim \mathrm{PY}(d, \alpha, H) \tag{2.35}$$

其中 $0 \leqslant d < 1$ 是折扣参数(discount parameter)，$\alpha > -d$ 是集中参数(concentrate parameter)，H 是基础测度。当 $d = 0$ 时，Pitman Yor 过程退化为带有集中参数 $\alpha > 0$ 的狄利克雷过程。Pitman-Yor 过程生成 G 的过程如下：

$$G = \sum_{i=1}^{\infty} \beta_i \delta_{\theta_i} \tag{2.36}$$

其中的 θ_i 由基础测度 H 独立同分布生成，而权重 β_i 生成如下：

$$\begin{gathered} \upsilon_i \mid d, \alpha \sim \mathrm{Beta}(1-d, \alpha+id),\ i = 1,2,\cdots \\ \beta_i = \upsilon_i \prod_{j=1}^{i-1}(1-\upsilon_j) \end{gathered} \tag{2.37}$$

无限序列 $(\beta_1, \beta_2, \cdots)$ 的联合分布为 $\mathrm{GEM}(d, \alpha)$。图 2.5 绘制了 (d, α) 分别为 $(0,5)$，$(0.1,5)$，$(0.5,5)$ 在 $k = 1,10,20$ 时的概率密度。

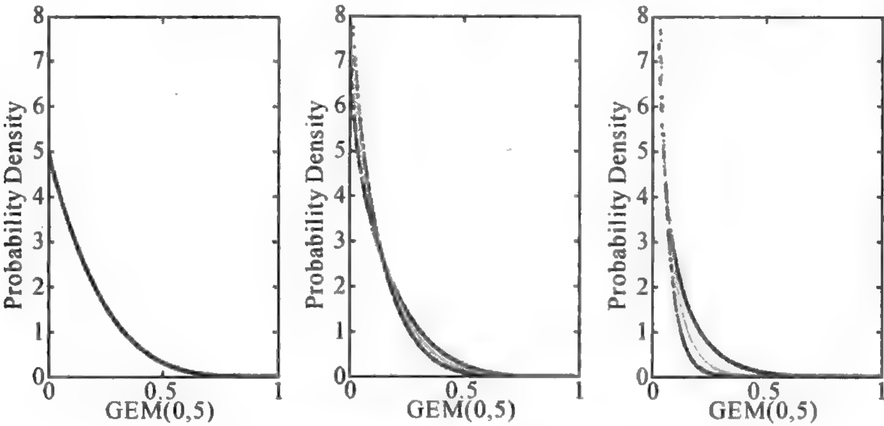


图 2.5　Pitman-Yor 过程中权重序列的联合分布

由式(2.36)和(2.37)可以看出，Pitman-Yor 过程的构造与 Stick-breaking 过程的构造非常相似，如果 Pitman-Yor 过程的折扣参数 $d = 0$ 即可得到 Stick-breaking 过程。但折扣参数影响着 Pitman-Yor 过程对棍子折断点的选取，折扣参数越大，每次选取的折断点越靠近上次折断的位置，这也意味着在相同条件下，带有折扣参数的 Pitman-Yor 过程比 Stick-breaking 过程获得的折断的棍子长度更均匀。

如果由同一 Pitman-Yor 过程得到 i 个参数，即 $\theta_1, \theta_2, \cdots, \theta_i \sim G, G = \mathrm{PY}(d, \alpha, H)$，假设已经生成前 $i-1$ 个参数，那么对于第 i 个参数的生成，可以通过对 G 积

分后得到θ_i的后验分布：

$$\theta_i \mid \theta_1, \theta_2, \cdots, \theta_{i-1}, d, \alpha, H \sim \sum_{t=1}^{K} \frac{n_t - d}{\alpha + i - 1} \delta_{\theta_t^*} + \frac{\alpha + Kd}{\alpha + i - 1} H \tag{2.38}$$

其中 K 是前 $i-1$ 个参数 $\theta_1, \theta_2, \cdots, \theta_{i-1}$ 中唯一值的个数，θ_t^* 是其中的第 t 个唯一值，n_t 是这个值在前 $i-1$ 个参数中出现的次数。这种描述与中国餐馆过程类似，θ_i 是第 i 个客人，θ_t^* 是餐馆中的第 t 张桌子。如果第 i 个客人坐在第 t 张桌子旁，则 $\theta_i = \theta_t^*$。与 Pitman-Yor 过程对 Stick-breaking 过程的影响类似，Pitman-Yor 过程增加了折扣参数对过程的影响。与中国餐馆过程以高概率产生大桌不同，折扣参数提高了新桌产生的概率，从而限制了大桌的产生。

2.5.4 狄利克雷构造过程与稀疏

在层次的 Stick-breaking 过程中，对层次模型(2.33)中 β_i, π_k 的特征值进行分析，β_i 与 π_k 的均值相同，即

$$\mathrm{E}[\pi_k] = \mathrm{E}[\beta_i] = \gamma^{k-1}(1+\gamma)^{-k} \tag{2.39}$$

π_k 的方差是 β_i 的方差的 $\mathrm{E}\left[\frac{\beta_i(1-\beta_i}{1+\alpha}\right]$ 倍，即

$$\mathrm{Var}[\pi_{ki}] = \mathrm{E}\left[\frac{\beta_i(1-\beta_i)}{1+\alpha}\right]\mathrm{Var}[\beta_i] \tag{2.40}$$

Teh 以柱形图的形式对 $\boldsymbol{\beta}$ 和 $\boldsymbol{\pi}$ 之间的关系进行了描述，如图 2.6 所示。由图可以看出，π_k 与 β_i 相比更具稀疏性，这种层次 Stick-breaking 过程引发的稀疏性在稀疏降维问题中有重要的意义。

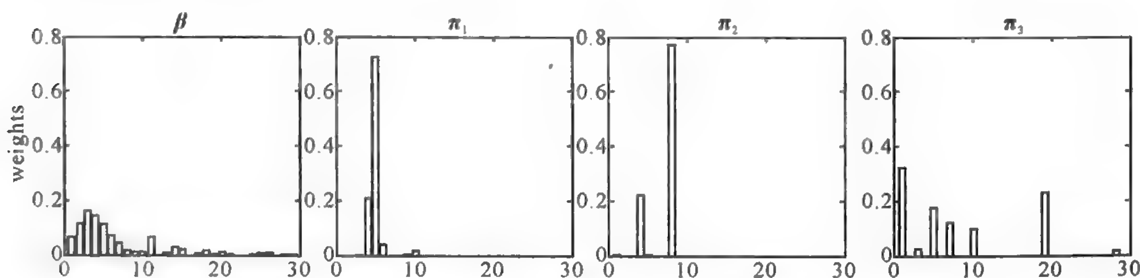


图 2.6　层次 Stick-breaking 过程

对于中国餐馆过程中顾客和餐桌的关系，如果把中国餐馆过程中一个顾客作为矩阵中的一行，一张桌子作为矩阵中的一列，则构成一个二元矩阵，顾客 i 坐在桌子 k 旁，则矩阵中 $(i,k)=1$，i 行中其余原子皆为 0。k 列中为 1 的原子个数与表示坐在该桌旁的顾客数。这种扩张给出了一种常见的假设，即对于每个对象，描述它是否具有某种特征，有则为 1，没有则为 0。把中国餐馆过程扩张为二元矩阵的形式对稀疏矩阵补充问题具有良好的表示。矩阵补充通常假设矩阵低秩或近似低

秩，在只有少量观察的情况下恢复矩阵的原始信息。基于狄利克雷过程的矩阵补充可以发挥贝叶斯非参数方法无限维度的优点，根据观测数据自适应恢复矩阵的信息，这也是狄利克雷过程的一个重要应用研究方向。

2.6 贝塔过程

2.6.1 贝塔过程的描述

贝塔过程(Beta Process)由 Hjort 在 1990 年提出并将其应用于基因分析。本书采用 Thibaux 和 Jordan 给出的定义：

定义 2.5：(贝塔过程，Beta Process)贝塔过程 $B \sim \text{Beta}(\alpha, B_0)$ 是一个正的 Lévy 过程，该过程的 Lévy 测度依赖于参数 α 和 B_0，其中，α 是 Θ 空间中的正函数，称为集中函数。当它是常量时，称之为集中参数；B_0 是 Θ 空间中的测度，称为基础测度。

如果基础测度 B_0 是连续的，则贝塔过程的 Lévy 测度是

$$\nu(\mathrm{d}\theta, \mathrm{d}\omega) = \alpha(\theta_i)\omega^{-1}(1-\omega)^{\alpha(\theta_i)-1}\mathrm{d}\omega B_0(\mathrm{d}\theta) \tag{2.41}$$

要得到 $B \sim \text{BP}(\alpha, B_0)$，首先根据基础测度 ν 由泊松过程在空间[0,1]获得一系列 ω_i，同时，在空间 Θ 获得一系列 θ_i，构成 (θ_i, ω_i)，令

$$B = \sum_{i=1}^{\infty} \omega_i \delta_{\theta_i} \tag{2.42}$$

即获得贝塔过程。这相当于是把空间 Θ 分为小的区域，将原子根据基础测度 B_0 和原子的权重(原子的权重由贝塔分布生成)，将其投入相应区域中，然后计算式(2.42)所示的和。

如果 B_0 是离散的，且 $B_0 = \sum_i q_i \delta_{\theta_i}, q_i \in [0,1]$，则 B 也由相同位置的原子构成：

$$B = \sum_i \omega_i \delta_{\theta_i} \tag{2.43}$$

其中

$$\omega_i = \text{Beta}(\alpha(\theta_i) q_i, \alpha(\theta_i)(1-q_i)) \tag{2.44}$$

图 2.7 描述了 $\alpha = 0$，$B_0 \sim \text{Uniform}(0,1)$ 的贝塔过程的采样，正如 Lévy 过程描述所示，贝塔过程得到的 B 是离散的，由于 $\nu(\Theta \times [0,1]) = \infty$，泊松过程生成无限多个点，使得式(2.43)的和由无穷多个采样点构成。

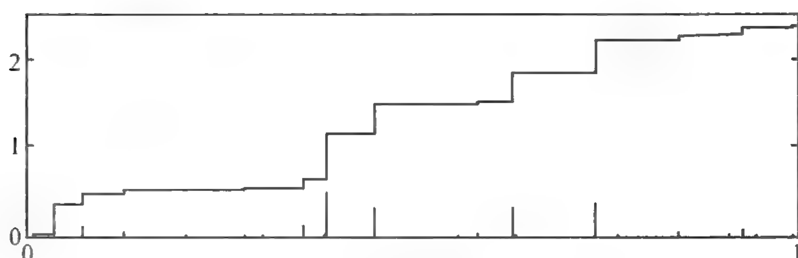


图 2.7 贝塔过程采样

2.6.2 贝塔过程的构造方法

与狄利克雷过程相同，贝塔过程也有多种构造方法，除了前面描述的 Lévy 过程，还有基于贝努利过程的构造、基于泊松过程的构造等方法。

1) 贝努利过程与贝塔过程

与贝塔过程共轭的是贝努利过程 $\mathrm{BeP}(B)$。贝努利过程采样结果只有 0 或 1 两种，原子只能出现在由基础测度 B 生成的位置，其是否出现由贝努利分布决定，且原子之间相互独立。由 n 次贝努利过程采样生成的二值矩阵为 $n\times\infty$ 矩阵，列数由基础测度决定。矩阵中大多数的元素都为 0，少数为 1，如图 2.8 所示。

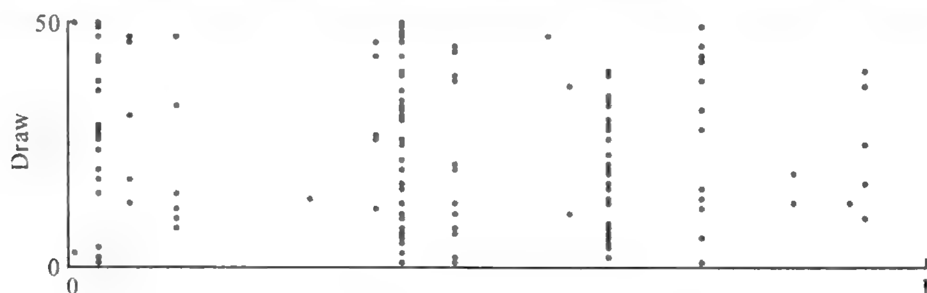


图 2.8 贝努利过程采样

基于贝努利过程和贝塔过程的共轭关系，可以得到如下的层次模型：

$$Z_i \mid B \sim \mathrm{BeP}(B),\ i=1,2,\cdots,n \tag{2.45}$$
$$B \mid \alpha, B_0 \sim \mathrm{BP}(\alpha, B_0)$$

如果 $Z_1, Z_2, \cdots, Z_n$ 在给定条件 B 下相互独立，则根据 De Finetti 定理，后验分布也是贝塔过程

$$B \mid Z_1, \cdots, Z_n, \alpha, B_0 \sim \mathrm{BP}\left(\alpha+n, \frac{\alpha}{\alpha+n}B_0+\frac{1}{\alpha+n}\sum_{i=1}^{n} Z_i\right) \tag{2.46}$$

2)基于泊松过程的构造

基于泊松过程的构造方法主要针对贝塔过程中在[0,1]和空间 Θ 上随机采样获得 ω_i 和 θ_i，获得的原子个数服从泊松过程。对第 $n(n\geqslant 1)$ 次服从贝塔过程的采

样过程如下：

首先根据泊松分布获得原子个数：

$$K_n \sim \text{Poisson}\left(\frac{\alpha\gamma}{\alpha + n - 1}\right) \tag{2.47}$$

由基础测度 $\frac{1}{\gamma}B_0$ 生成 K_n 个 θ_i，再由 $\text{Beta}(1, \alpha + n - 1)$ 生成 ω_i；

更新：

$$\hat{B}_n = \hat{B}_{n-1} + \sum_{i=1}^{K_n} \omega_i \delta_{\theta_i} \tag{2.48}$$

这种构造方法与用 Stick-breaking 过程构造狄利克雷过程很类似，且依赖原子在 Θ 中的比重进行采样。

2.6.3 贝塔过程与稀疏

贝塔过程是一个离散的过程，每个原子的权重 $0 < \omega_i < 1$，且 $\sum_i \omega_i$ 可以不等于 1，这使得贝塔过程与狄利克雷过程相比，约束更为松弛。贝塔构建过程中无论是基于贝努利过程还是基于泊松过程，都在空间 Θ 中得到带有权重的原子点 δ_{θ_i}，而且其原子点的权重为 0，自然可以用于稀疏问题的建模。

本书的第三章和第四章中充分利用贝塔过程对稀疏表示、字典构建过程进行建模，并与其他贝叶斯非参数方法进行比较，实验证实了贝塔过程对稀疏问题的表示的优势。

2.7 小　结

本章在介绍贝叶斯非参数方法的基本理论的基础上，阐明了狄利克雷分布向无限维度扩展生成的狄利克雷过程的方法，分析了狄利克雷过程构建方法和贝塔过程构建方法对稀疏建模的可能性和优势，给出了基于不同构建方法实现稀疏表示的策略。

第3章 贝叶斯稀疏表示

稀疏表示作为一种重要的数据编码与表达方式，不仅在人类的视觉认知机理上具有明确的理论依据，而且在信号表达与重建理论方面得到了严格的证明和推导。神经生理学机制已经揭示了稀疏表达作为一种广泛的视觉先验，在视觉认知和推理过程中发挥着重要作用。压缩感知(Compressive Sensing，CS)理论从信号表达的角度证明了稀疏表达是高维信号在特定基向量或者"字典(Dictionary)"上的一种自然表达，由此发展的约束优化求解策略为信号的稀疏表达提供了近似最优的可计算模型。目前，稀疏表达已经在理论和方法上得到了快速的发展，并在信号压缩、图像处理、模式识别、机器学习等多个应用领域取得了很多成功的应用；同时由于稀疏表达在视觉认知上的理论基础，该方法对视觉任务的应用具有很多天然的优势。

本章首先对稀疏表示问题进行描述，介绍常见的求解方法，然后针对已有贝叶斯稀疏表示求解方法进行分析，再利用贝塔过程将贝叶斯稀疏表示问题中的先验模型从收敛先验扩展到离散混合先验，使其能够根据观测数据自适应调整稀疏信号的稀疏度，最后通过实验进行验证。

3.1 稀疏表示

3.1.1 问题描述

首先考虑线性模型

$$y = X\beta \tag{3.1}$$

其中 y 是 $N\times 1$ 维采样信号，X 是 $N\times K$ 维矩阵，$N\ll K$，β 是 $K\times 1$ 维系数向量。给定 X 和 y，式(3.1)是一个欠定问题，有无穷多组解。如果要求 β 是稀疏的，即向

量 $\boldsymbol{X}$ 中非零项的个数尽可能小，则问题可描述为：

$$\arg\min \|\boldsymbol{\beta}\|_0 \quad \text{s.t.} \quad \boldsymbol{X\beta} = \boldsymbol{y} \tag{3.2}$$

其中，$\|\boldsymbol{\beta}\|_0$ 是 l_0 范数，$\|\boldsymbol{\beta}\|_0 = \#\{j, \beta_j \neq 0\}$。

Donoho 等人证明，如果矩阵 $\boldsymbol{X}$ 满足 $\sigma(\boldsymbol{X}) \geqslant 2\|\boldsymbol{\beta}\|_0$，则 l_0 范数优化问题具有唯一的解，其中 $\sigma(\boldsymbol{X})$ 是最小的线性相关的列向量集所含的向量个数。但 Donoho 也指出，最小 l_0 范数问题是一个 NP-hard 问题，需要穷举 $\boldsymbol{\beta}$ 中非零值的所有 C_K^N 种排列可能。

2006 年，Terrence Tao 与 Candès 合作证明了在满足约束等距性条件（RIP）下，l_0 范数优化问题与以下 l_1 范数优化问题具有相同的解：

$$\arg\min \|\boldsymbol{\beta}\|_1 \quad \text{s.t.} \quad \boldsymbol{X\beta} = \boldsymbol{y} \tag{3.3}$$

其中约束等距性条件为：存在满足某种条件的常数 μ_K，使得

$$(1-\mu_K)\|\boldsymbol{\beta}\|_2^2 \leqslant \|\boldsymbol{X\beta}\|_2^2 \leqslant (1+\mu_K)\|\boldsymbol{\beta}\|_2^2, \forall \boldsymbol{\beta}, \|\boldsymbol{\beta}\|_0 \leqslant K \tag{3.4}$$

l_1 范数优化问题是一个凸优化问题，存在唯一解。

进一步考虑含噪声的情况，可得到相似的结果。

$$\arg\min \|\boldsymbol{\beta}\|_0 \quad \text{s.t.} \quad \|\boldsymbol{X\beta} - \boldsymbol{y}\|_2^2 \leqslant \varepsilon \tag{3.5}$$

其中 ε 是一个微小数值。

在上述数学证明基础之上，稀疏化问题的求解方法层出不穷，大致可分为三类：①直接优化 l_0 范数的贪婪算法。②使用 l_p 范数近似计算 l_0 范数的凸优化方法。③以稀疏贝叶斯为代表的统计优化算法。下面简单介绍贪婪算法和 l_p 范数近似计算 l_0 范数的凸优化方法，贝叶斯方法的稀疏表示作为本文的研究重点，将在下一节讨论。

3.1.2 贪婪算法

贪婪算法是针对组合优化提出的，代表算法有匹配追踪（Matching Pursuit）、正交匹配追踪（Orthogonal Matching Pursuit）等。

基于组合优化的方法求新信号稀疏表示的目标，是在已知的字典 $\boldsymbol{X}$ 中选出一个包含 M 个向量的子集 $\boldsymbol{x}_{r1}, \cdots, \boldsymbol{x}_{rM}$，使得在同样适用 M 项来逼近信号的情况下，误差最小，即

$$(\boldsymbol{\beta}, \boldsymbol{X}) = \arg\min \|\boldsymbol{y} - \sum_{m=1}^{M} x_{rm}\beta_{rm}\|^2 \tag{3.6}$$

满足式(3.6)的逼近称为信号 $\boldsymbol{y}$ 的 M 项最优逼近。但直接根据式(3.6)来求解信号的 M 项最优逼近是 NP 难题。Mallt 提出的匹配追踪算法将这个最优逼近问题通过迭代的贪婪算法，化简为求 M 个单项最优逼近的问题。在每次迭代过程中，从矩阵 $\boldsymbol{X}$ 中选择最能匹配信号的一个列向量，从而构成逐步近似求解信号的

稀疏化表达过程。

算法首先按照某种规则选定向量 $\boldsymbol{x}_{r1} \in \boldsymbol{X}$，将 $\boldsymbol{y}$ 分解成沿着 $\boldsymbol{x}_{r1}$ 方向的分量和与其垂直方向的分量的叠加：

$$\boldsymbol{y} = \langle \boldsymbol{y}, \boldsymbol{x}_{r1} \rangle \boldsymbol{x}_{r1} + R\boldsymbol{y} \tag{3.7}$$

则其中 $R\boldsymbol{y}$ 是信号 $\boldsymbol{y}$ 沿 $\boldsymbol{x}_{r1}$ 分解后的残差。根据勾股定理，有

$$\| \boldsymbol{y} \|^2 = | \langle \boldsymbol{y}, \boldsymbol{x}_{r1} \rangle |^2 + \| R\boldsymbol{y} \|^2 \tag{3.8}$$

为了使 $\| R\boldsymbol{y} \|$ 最小，应选取使 $| \langle \boldsymbol{y}, \boldsymbol{x}_{r1} \rangle |$ 尽可能大的 x_{r1}，即

$$| \langle \boldsymbol{y}, \boldsymbol{x}_{r1} \rangle | = \sup_{r \in I} | \langle \boldsymbol{y}, \boldsymbol{x}_{r1} \rangle | \tag{3.9}$$

再对残差 $R\boldsymbol{y}$ 同样在字典 $\boldsymbol{X}$ 中找到与之最匹配的向量，再次得到残差；重复这个过程直到残差小于预设的值。

经过 M 次迭代，信号 $\boldsymbol{y}$ 被分解为：

$$\boldsymbol{y} = \sum_{m-1}^{M} \langle R^m \boldsymbol{y}, \boldsymbol{x}_{rm} \rangle \boldsymbol{x}_{rm} + R^M \boldsymbol{y} \tag{3.10}$$

匹配追踪算法是目前最广泛应用的求解稀疏表示的方法，这种近似方法得到的稀疏度虽然不够高，但计算复杂度大大降低。然而该算法的一个明显缺点是，在已选列向量组成的子空间上，它不是一个正交投影，因此信号的展开可能不是最优的。

正交匹配追踪(OMP)算法是在匹配追踪算法基础上的一种改进算法，此算法选取最佳列向量的方法与匹配追踪算法一样，不同的是正交匹配追踪算法将所选列向量利用 Gram-Schmidt 正交化方法进行正交化处理，再将信号在这些正交列向量构成的空间上投影，得到信号在各个已选列向量上的分量和残余分量；然后用与匹配追踪相同方法分解残余分量。经过 M 次分解，原始信号被分解为 M 个原子的线性组合。在每一步分解中，所选最佳列向量均满足一定条件，因此，残余分量随着分解迅速减少，这样，用少量列向量就可以表示原始信号，而经过有限次迭代就可以收敛。

3.1.3 凸优化方法

用凸优化方法 l_p 范数近似求解 l_0 范数，主要有 Lasso、岭回归、弹性网等方法。

Lasso 方法采用 l_1 范数约束代替 l_0 范数约束求解稀疏表示，由于 Lasso 方法用回归模型系数的绝对值函数作为惩罚来压缩模型系数，使得绝对值较小的系数自动为零，从而实现模型参数选择的自然稀疏性。与传统的模型选择方法相比，Lasso 方法很好地克服了传统方法在选择模型上的不足，因此该方法在统计领域受到了极大的重视。在算法方面，最初用二次规划方法做 Lasso 回归，但其有效性不能满足人们的需求，因此，很多学者在这方面展开研究，包括 Fu 提出了

"Shooting"算法、Osborne 等提出了相应的同伦算法等。2002 年，Efron 等人提出的最小角回归(Least Angle Regression)算法很好地解决了 Lasso 的计算问题，该方法的计算复杂度与最小二乘回归相当。有效算法的提出使 Lasso 方法广为流行。

稀疏表达的过程可以通过优化一个"损失+惩罚"的函数问题来完成，这种方法一般被称为正则化方法。岭回归是 l_2 正则化方法。岭回归尽管可以有效克服自变量间的高度相关性，并能提高预测精度，但单纯用此方法却不能得到稀疏解。在岭回归的基础上引入"Boosting"后可以得到与 Lasso 同样的估计。

基于最小角回归算法的 Lasso 方法尽管有着非常好的性质，并且也的确克服了传统方法的一些不足，但是单纯针对 Lasso 运用最小角回归算法，对于 $K \gg N$ 的情形，最多只能选择 N 个自变量，往往得到过于稀疏的模型。针对这个问题，Zou 和 Hastie 提出一种处理该问题相当有效的方法——弹性网(elastic net)。

弹性网方法同时采用 l_1 范数和 l_2 范数约束，实现对 Lasso 方法的凸松弛，从而得到较"温和"的稀疏模型；当弹性网方法中 l_2 范数惩罚项的系数为零时，其退化为 Lasso 方法。在一些面向应用的特定稀疏建模中，不仅对模型系数有稀疏性要求，同时还要求为非负，比如图像像素值的生成，则可以在 Lasso 方法或者弹性网方法的基础上，增加对模型系数的非负约束。

3.2 贝叶斯稀疏表示方法

贝叶斯学习机制是将先验分布中的期望值与样本均值按各自的精度进行加权平均，精度越高者其权值越大。在先验分布为共轭分布的前提下，可以将后验信息作为新一轮计算的先验，用贝叶斯定理与进一步得到的样本信息进行综合。多次重复这个过程后，样本信息的影响越来越显著。由于贝叶斯方法可以综合先验信息和后验信息，既可避免只使用先验信息可能带来的主观偏见和缺乏样本信息时的大量盲目搜索与计算，也可避免只使用后验信息带来的噪声影响。

系统的采样过程中不可避免地包含噪声，所以，在下文中，均考虑含噪的线性模型

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{3.11}$$

其中 $\boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2 \boldsymbol{I}_n)$ 的高斯噪声。将式(3.11)写为高斯模型，得到：

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma_\varepsilon^2) = N(\boldsymbol{X\beta}, \sigma_\varepsilon^2 \boldsymbol{I}) \tag{3.12}$$

合理的先验假设，是贝叶斯方法进行有效学习的关键。对于稀疏表示问题，根据信号 β 的稀疏特性，直观采用先验是 Laplace 先验，但由于 Laplace 先验与似然

函数是非共轭的，所以在计算中不易处理。因此，研究者研究和发展了多种基于不同先验分布的贝叶斯稀疏表示学习方法，主要有相关向量机、基于高斯先验的稀疏表示、贝叶斯等。

3.2.1 相关向量机

相关向量机（RelevanceVector Machine，RVM）是在贝叶斯框架下进行学习的稀疏概率模型。RVM在先验参数的结构下基于主动相关决策理论（Automatic Relevance Determination，ARD）来移除不相关的点，从而获得稀疏化的模型。

对于给定的训练样本 $\{\boldsymbol{y}_i,\boldsymbol{x}_i\}_{i=1}^{N}$，RVM的模型输出定义为：

$$\begin{aligned} p(\boldsymbol{y}_i) &= N(\boldsymbol{y}_i \mid f(\boldsymbol{x}_i;\boldsymbol{\beta}),\sigma^2) \\ f(\boldsymbol{x}_i;\boldsymbol{\beta}) &= \sum_{k=1}^{K}\beta_k K(\boldsymbol{x},\boldsymbol{x}_i)+\boldsymbol{\beta}_0 \end{aligned} \tag{3.13}$$

其中 $K(\cdot)$ 表示核函数。

假设 $\{\boldsymbol{y}_i\}_{i=1}^{N}$ 是彼此独立的，则RVM模型得到 $\boldsymbol{y}$ 的概率分布为：

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\beta},\sigma^2) &= \prod_{i=1}^{M} N(\boldsymbol{y}_i \mid f(\boldsymbol{x}_i;\boldsymbol{\beta}),\sigma^2) \\ &= (2\pi\sigma^2)^{-2/N}\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{\Phi\beta}\|^2}{2\sigma^2}\right) \end{aligned} \tag{3.14}$$

其中 $\boldsymbol{\beta}$ 是由 β_k 组成的向量，$\boldsymbol{\Phi}$ 则是由各向量 $\boldsymbol{x}_i$ 输入核函数得到的 $N\times(N+1)$ 矩阵：

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & K(\boldsymbol{x}_1,\boldsymbol{x}_1) & K(\boldsymbol{x}_1,\boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_1,\boldsymbol{x}_N) \\ 1 & K(\boldsymbol{x}_2,\boldsymbol{x}_1) & K(\boldsymbol{x}_2,\boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_2,\boldsymbol{x}_N) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & K(\boldsymbol{x}_N,\boldsymbol{x}_1) & K(\boldsymbol{x}_N,\boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_N,\boldsymbol{x}_N) \end{bmatrix}$$

相关向量机为每个权重 β_k 定义了一个独立的零均值高斯先验的概率分布：

$$p(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) = \prod_{k=0}^{K} N(\beta_k \mid 0,\alpha_k^{-1}) \tag{3.15}$$

其中 α_k 表示高斯密度函数的精确度（是方差的倒数），它的先验为：

$$p(\boldsymbol{\alpha} \mid a,b) = \prod_{k=1}^{K}\Gamma(\alpha_k \mid a,b) = \prod_{k=1}^{K}\frac{b^a}{\Gamma(a)}\alpha_k^{a-1}\exp(-ba) \tag{3.16}$$

通过对超参数 α 进行边缘分布计算，得到 β 的先验为

$$p(\boldsymbol{\beta} \mid a,b) = \prod_{k=1}^{K}\int_0^{\infty} N(\beta_k \mid 0,\alpha_k^{-1})\Gamma(\alpha_k \mid a,b)\mathrm{d}\alpha_k \tag{3.17}$$

其中$\int_0^{\infty} N(\beta_k \mid 0, \alpha_k^{-1}) \Gamma(\alpha_k \mid a, b) \mathrm{d}\alpha_k$符合 Student-t 分布，选择合适的 a 和 b 时，Student-t 分布在 $\beta_k = 0$ 附近获得峰值，因此这个先验促进 β 的稀疏。类似的，可以为 σ^2 选择 $\Gamma(\alpha_0 \mid c, d)$ 先验。RVM 图模型如图 3.1 所示。

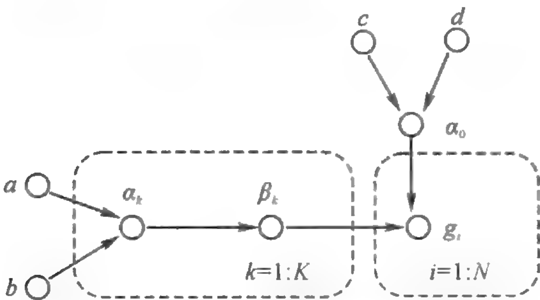


图 3.1　RVM 图模型

由于 RVM 具有很好的稀疏性及核函数的任意性，RVM 应用到越来越多的领域，比如医疗诊断、图像处理、视觉跟踪、时间序列预测等。但是计算的复杂性和占用大量的存储空间限制了 RVM 应用到大数据集中。目前解决这个困难的方法有快速边际似然法、将核函数正交分解的 Gram-Schmidt 算法、Boosting RVM 等。

3.2.2 基于高斯先验的稀疏表示

基于高斯先验的稀疏表示模型有基于 Normal-Jeffreys，Normal-Gamma 和 Normal-Inverse Gaussian 等模型。在这些模型中，假设 $\{\beta_k\}_{k=1}^{K}$ 之间是独立的，且服从均值为 0，方差为 σ_k^2 的高斯分布：

$$p(\beta_k) = \int N(\beta_k; 0, \sigma_k^2) p(\sigma_k^2) \mathrm{d}\sigma_k^2 \tag{3.18}$$

对于方差 σ_k^2，Normal-Jeffreys 采用先验 $p(\sigma_k^2) \propto 1/\sigma_k^2$。

计算所得后验为：

$$p(\boldsymbol{\beta}) \propto -\boldsymbol{\beta}^T H^T H \boldsymbol{\beta} - 2\boldsymbol{\beta} H^T \boldsymbol{y} - \boldsymbol{\beta}^T \Upsilon(\sigma_\beta^2) \boldsymbol{\beta} \tag{3.19}$$

其中 H 是 $\boldsymbol{x}$ 的核函数，$\Upsilon(\sigma_\beta^2) = \mathrm{diag}(\sigma_1^{-2}, \cdots, \sigma_K^{-2})$。这种先验不是标准的先验，因为它的积分不是有限的，这也被称为不恰当的先验（improper prior），这种先验也不能够生成对 $\boldsymbol{\beta}$ 的 Laplace 先验，但 Figueiredo 的实验表明，该先验能够有效引起稀疏，并且效果良好。

Normal-Gamma 和 Normal-Inverse Gaussian 分别用 Gamma 函数和 Inverse-Gaussian 函数作为 σ_k^2 的先验，即 $\sigma_k^2 \sim \mathrm{Gamma}(\frac{\alpha}{K}, \frac{\gamma^2}{2})$ 和 $\sigma_k^2 \sim \mathrm{IG}(\frac{\alpha}{K}, \gamma)$，得到的后验为

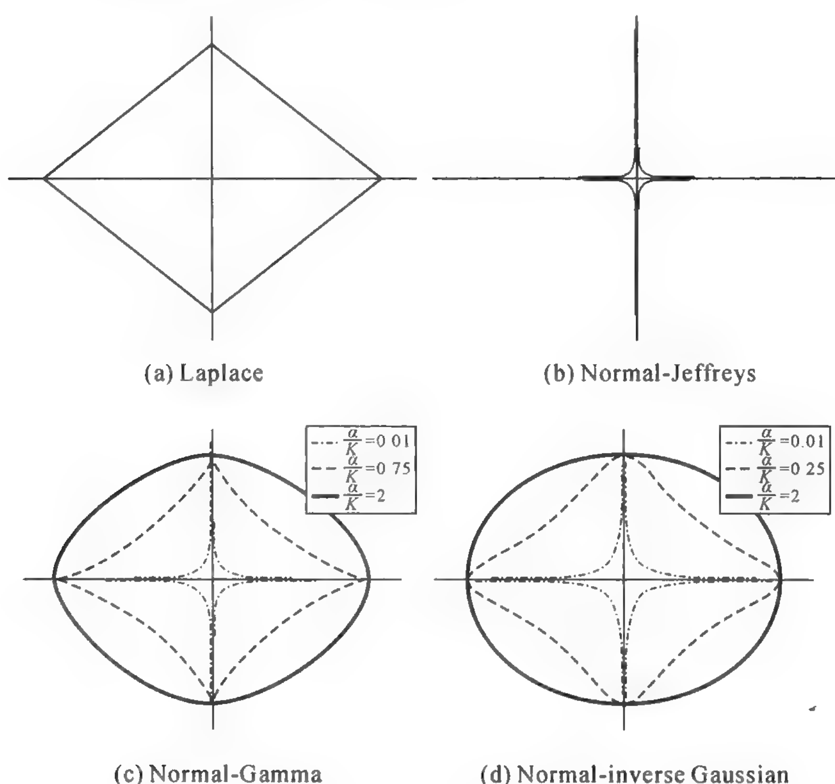
$$p(\boldsymbol{\beta}) \propto |\beta_k|^{\frac{\alpha}{K}-\frac{1}{2}} K_{\frac{\alpha}{K}-\frac{1}{2}}(\gamma|\beta_k|) \tag{3.20}$$

和

$$p(\boldsymbol{\beta}) \propto \left(\frac{\alpha^2}{K^2}+\beta_K^2\right)^{-1/2} K_1\left(\gamma\sqrt{\frac{\alpha^2}{K^2}+\beta_K^2}\right) \tag{3.21}$$

Normal-Gamma 先验能够引发稀疏估计，而 Normal-Inverse Gaussian 先验能够引发近似稀疏的估计，即大多数的协方差都趋向于 0。

图 3.2 显示了不同先验的轮廓，当 $\alpha/K = 1$ 时，Normal-Gamma 先验等于 Laplace 先验，则问题退化为 l_1 惩罚，可以采用 Lasso 方法。当 $\alpha/K \to 0, c \to 0$，则先验是 Normal-Jeffreys，惩罚退化为 $\log(|\beta_k|)$。



(a) Laplace　(b) Normal-Jeffreys

(c) Normal-Gamma　(d) Normal-inverse Gaussian

图 3.2　不同先验的轮廓

综合上述三种基于高斯先验的稀疏模型，可以用下式表示：

$$-\sum_{i=1}^{N}\frac{1}{2\sigma_i^2}\|\boldsymbol{y}_i-\boldsymbol{X}\beta_i\|^2-\sum_{k=1}^{K}\text{pen}(\beta_k^{1:N}) \tag{3.22}$$

其中 Normal-Jeffreys，Normal-Gamma 和 Normal-Inverse Gaussian 模型中 $\text{pen}(\beta_k^{1:N})$ 如表 3.1 所示。其中 $\mu_k = \sqrt{\sum_{i=1}^{N}(\beta_k^i)^2}$，$q_k = \sqrt{\frac{\alpha^2}{K^2}+\mu_k^2}$。

表 3.1 基于高斯先验的三种模型

模型	$\mathrm{pen}(\beta_k^{1:N})$
Normal-Jeffreys	$N\log(\mu_k)$
Normal-Gamma	$\left(\frac{N}{2}-\frac{a}{K}\right)N\log(\mu_k)-\log K_{\frac{a}{K}-\frac{1}{2}}(\gamma\mu_k)$
Normal-Inverse Gaussian	$\frac{N+1}{2}\log(q_k)-\log K_{\frac{N+1}{2}}(\gamma q_k)$

3.2.3 贝叶斯 l_p 范数

贝叶斯 l_p 范数有针对 l_1 范数构成的贝叶斯 Lasso、针对 l_1 和 l_2 范数构成的贝叶斯弹性网等方法，主要采用适合的先验，通过积分所得后验得到与 l_p 范数相似的形式，从而实现贝叶斯 l_p 范数。

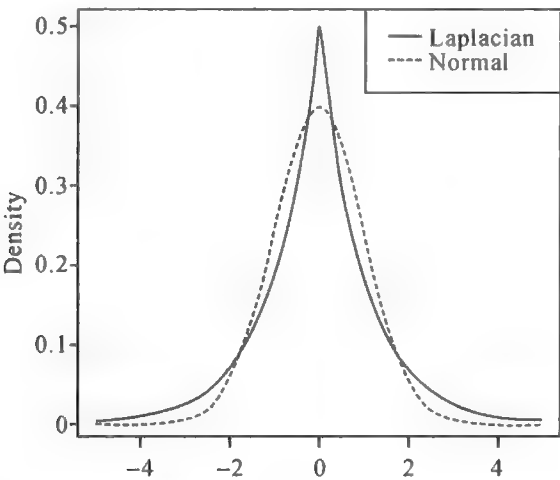


图 3.3 Laplace 分布

贝叶斯 Lasso 基于 Laplace 先验进行构造。Laplace 先验是稀疏表示的最直观的先验形式，而且与正态分布相比，Laplace 分布更集中于 0，如图 3.3 所示。但 Laplace 先验与高斯是非共轭的，这带来计算的复杂性，因此，Park 等人采用层次先验的形式：

$$\begin{gathered}\beta_k \mid \sigma^2,\gamma_k^2 \sim N(0,\sigma^2\gamma_k^2)\\ \gamma_k^2 \mid \sigma^2 \sim \exp(\lambda^2/2)\end{gathered} \tag{3.23}$$

得到：

$$p(\beta_k \mid \sigma^2) \sim \frac{\lambda}{2\sqrt{\sigma^2}}\exp(-\lambda \mid \beta_k \mid / \sqrt{\sigma^2}) \tag{3.24}$$

对参数 λ，一种方法是通过对最大似然取边缘分布，再通过 EM 算法根据式

(3.25)对 λ 进行更新：

$$\lambda^{(k)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} \mathrm{E}_{\lambda^{(k-1)}}[\gamma_i^2 \mid \boldsymbol{y}]}} \tag{3.25}$$

另一种方法是为 λ^2 设置在 λ 的边缘最大似然估计值附近取得高概率的超先验。

在上述先验的基础上，根据贝叶斯 Lasso 公式，计算得到后验如下：

$$\begin{aligned} \boldsymbol{\beta} &\sim N(A^{-1}\boldsymbol{X}^T\boldsymbol{y}, \sigma^2 A^{-1}) \\ \sigma^2 &\sim \mathrm{InvGamma}(a, b) \\ 1/\gamma_i^2 &\sim \mathrm{InvGaussian}(a_0, b_0) \end{aligned} \tag{3.26}$$

其中，$\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{D}_\gamma^{-1}$，$\boldsymbol{D}_\gamma = \mathrm{diag}(\gamma_1^2, \cdots, \gamma_p^2)$；$\sigma$ 的后验分布中，参数 $a = (n+p)/2$，$b = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})/2 + \boldsymbol{\beta}^{-1} D_\gamma^{-1}\boldsymbol{\beta}/2$，参数 $a_0 = \sqrt{\lambda^2\sigma^2/\beta_j^2}$，$b_0 = \lambda^2$。

贝叶斯弹性网基于弹性网的 l_1 范数和 l_2 范数约束，得到如下模型：

$$\begin{aligned} \beta_k \mid \sigma^2, \alpha_i, \lambda &\sim N(0, \sigma^2(\alpha_k + \lambda)) \\ \alpha_k &\sim \eta\left(\frac{\alpha_k}{\alpha_k + \lambda}\right)^{1/2} \mathrm{IG}(\alpha_k; 1, \frac{\gamma}{2}) \end{aligned} \tag{3.27}$$

对 α 积分之后，得到似然函数：

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \gamma) \propto f(\sigma^2, \gamma)\exp\left\{-\frac{1}{2\sigma^2}(\| \boldsymbol{y} - \boldsymbol{X\beta} \|^2 + 2\sqrt{\gamma\sigma^2} \mid \boldsymbol{\beta} \mid + \lambda \| \boldsymbol{\beta} \|^2)\right\}$$

其中，$f(\sigma^2, \gamma)$ 是 σ^2 和 γ 的联合分布。这个似然函数与弹性网模型相似，得到 β 的后验为：

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\mathrm{argmin}}\{\| \boldsymbol{y} - \boldsymbol{X\beta} \|^2 + 2\sqrt{\gamma\sigma^2} \mid \boldsymbol{\beta} \mid + \lambda \| \boldsymbol{\beta} \|^2\} \tag{3.28}$$

通过为 γ_k 选择合适的 Gamma 先验得到稀疏 β。

将 l_p 范数惩罚先验嵌入在层次贝叶斯中有很多优点，除了常说的层次模型容易解释之外，贝叶斯公式为不确定性提供了可用的测度。然而，上面介绍的求解稀疏表示的方法均假设信号的稀疏度 S 是已知的，然而在许多情况下，S 并非事先已知，需要根据观测数据估计 $\boldsymbol{\beta}$ 稀疏度，因此建立动态的测量方式和相应的重建算法是关键的问题，而贝叶斯非参数方法正是适应于建立这种动态关系的有效方法。

3.2.4 贝叶斯非参数稀疏表示

对于稀疏度未知条件下的贝叶斯 l_p 范数的模型已有若干研究，基于贝叶斯非参数过程的稀疏表示模型展现了良好的模型自适应性和稀疏表示能力。

Maclehose 和 Dunson 针对观测值 $\boldsymbol{y}$ 是二值数据的情况，采用 Lasso 先验 $p(\boldsymbol{\beta}) = \prod_{k=1}^{K} \mathrm{DE}(\beta_k \mid 0, \tau)$ 对 $\boldsymbol{\beta}$ 进行约束，其中 0 是位置参数，τ 是范围参数。根

据 West 对范围混合高斯的表示，有：

$$\mathrm{DE}(\beta_k \mid 0,\tau) = \int_0^{\infty} N(\beta_k \mid 0,\lambda_k)\exp(\lambda_k \mid 2/\tau)\mathrm{d}\lambda_k \tag{3.29}$$

其中 $\exp(\lambda_k \mid 2/\tau)$ 是以 $2/\tau$ 为均值的指数分布。由于模型的观测值 $\boldsymbol{y}$ 是二值的，所以式(3.29)中的条件分布是共轭的。在上述分析的基础上，Maclehose 和 Dunson 对于收敛到非 0 的 $\boldsymbol{\beta}$ 给出先验约束 $p(\boldsymbol{\beta}) = \prod_{k=1}^{K}\mathrm{DE}(\beta_k \mid \mu_k,\lambda_k)$，再通过狄利克雷过程分别为参数赋予层次先验，得到模型如式(3.30)所示。

$$\begin{aligned}
\beta_k &\sim N(\beta_k \mid \mu_k,\lambda_k) \\
\lambda_k &\sim \exp(\lambda_k \mid 2/\tau) \\
(\mu_k,\tau_k) &\sim \pi\delta_o(\mu_k)\mathrm{Gamma}(\tau_k \mid a_0,b_0) + (1-\pi)D \\
\pi &\sim \mathrm{Beta}(\pi \mid 1,\alpha) \\
D &\sim \mathrm{DP}(\alpha D_0) \\
D_0 &= N(\mu_k \mid c,d)\mathrm{Gamma}(\tau_k \mid a_1,b_1)
\end{aligned} \tag{3.30}$$

该模型通过 Gibbs 采样进行计算。模型适应于观测数据是二值数据的情形，而且对于小样本数据有较好的处理能力，当样本数 N 或字典维度 K 较大时，计算速度急剧下降。

在基于贝叶斯非参数的稀疏表示的研究中，目前的热点集中于以贝塔过程对稀疏进行建模表示。Paisley 在 2009 年 ICML 会议上发表"*Nonparametric factor analysis with Beta process priors*"一文，继而基于 Stick-breaking 构造贝塔过程的方法、贝塔过程用于字典学习等方面开展了深入研究。在他们提出的模型中，$\boldsymbol{\beta}$ 被分解为二值因子 z 和权重因子 $\boldsymbol{\omega}$ 的 Hadamard 乘积，即将模型(3.11)转变为：

$$\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{z} \circ \boldsymbol{\omega}) + \boldsymbol{\varepsilon} \tag{3.31}$$

的形式，再对新的模型赋予先验，其中为 z 赋予贝努利-贝塔过程先验，对 ω 赋予高斯先验，构建的贝叶斯非参数层次模型 BP-FA 如式(3.32)所示：

$$\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X}(\boldsymbol{z} \circ \boldsymbol{\omega}) + \boldsymbol{\varepsilon} \\
\boldsymbol{\omega} &\sim N(0,\sigma_\omega^2 I) \\
z_k &\sim \mathrm{Bernoulli}(\pi_k) \\
\pi_k &\sim \mathrm{Beta}(a/K, b(K-1)/K) \\
\boldsymbol{\varepsilon} &\sim N(0,\sigma_n^2 I)
\end{aligned} \tag{3.32}$$

其图模型更清楚直观地表示各参数之间的关系，如图 3.4 所示。

与直接对 $\boldsymbol{\beta}$ 建模为均值为 0 的高斯分布相比，BP-FA 模型通过参数 $\boldsymbol{\pi}$ 强化了在相同因子子集上的稀疏性。但模型对权重参数和稀疏参数的同步更新影响了模型的运算速度，因为既然通过贝努利-贝塔过程增加了在相同子集上得到一

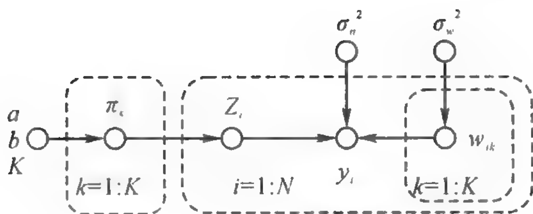


图 3.4　BP-FA 图模型

致稀疏值的概率，其权重的不断更新引起额外的、不必要的计算。当样本数 N 或字典维度 K 较大时，计算速度急剧下降。

3.3 基于离散混合贝塔过程的稀疏表示模型

在分析了已有基于贝叶斯方法的稀疏表示模型基础上，本节提出离散混合贝塔过程贝叶斯非参数模型（Discrete Mixture Prior Beta Process model, DMPBP）。模型首先通过离散混合模型将数据分为稀疏部分和非稀疏部分，再对非稀疏部分的参数进行估计，从而达到在计算过程中首先降低数据的维度，提高计算速度的目的。对离散混合因子，模型以贝努利-贝塔过程作为先验，利用贝努利-贝塔过程构造相对简单的特点快速得到稀疏因子；对于非稀疏部分，模型采用层次表示的 Laplace 分布作为先验，一方面利用 Laplace 分布比正态分布相比更集中于均值的特点进一步逼近稀疏，另一方面采用层次表示的方式降低 Laplace 先验与式(3.12)非共轭带来的计算的复杂性。

3.3.1 模型描述

离散混合先验（Discrete Mixture Priors）是元线性回归、小波滤波阈值、变量选择等问题中常见的先验模型，此模型的优点在于能够快速获得变量的稀疏性并迅速收敛。对于未知变量 $\boldsymbol{\beta}$ 中的每个分量，假设其先验为：

$$p(\beta_i \mid \omega_i, \gamma) = (1-\omega_i)\delta_0(\beta_i) + \omega_i\gamma(\beta_i) \tag{3.33}$$

其中 $\delta_0(\beta_i)$ 定义为，当 $\beta_i = 0$ 概率为 1，其他情况概率为 0 的函数。ω_i 是混合参数，其值为 0 或 1。如果 ω_i 为 0，则该 $\beta_i = 0$。$\boldsymbol{\omega}$ 控制着 $\boldsymbol{\beta}$ 的稀疏度 $(S = K - |\boldsymbol{\omega}| = K - \sum_i \omega_i)$。

$\gamma(\beta_i)$ 是 $\beta_i \neq 0$ 时收敛的函数。常见的收敛函数有正态概率分布函数 $N(0, \tau^2)$ 和 Laplace 分布函数。相对于正态概率分布函数，Laplace 分布函数的尾部概率更重，从而对于大数据问题有更好的适应性。通过这两种常见的收敛函数的形

式可以看出，模型期望非稀疏部分尽可能为 0，从而增加稀疏度。

在离散混合先验模型的基础上，本章针对稀疏表示问题提出如下模型：

$$\begin{aligned} \boldsymbol{y} &\sim N(\boldsymbol{X\beta}, \sigma^2 I) \\ \beta_k &\sim (1-\omega_k)\delta_0(\beta_k) + \omega_k f(\beta_k) \\ f(\beta_k) &\sim N(0, \sigma^2\gamma_k) \\ \gamma_k &\sim \exp(\tau^2/2) \\ \sigma^2 &\sim \text{Gamma}(a_0, b_0) \end{aligned} \tag{3.34}$$

其中 $f(\beta_k) \sim N(0,\sigma^2\gamma_k^2)$，$\gamma_k \sim \exp(\tau^2/2)$ 是 Laplace 先验的层次表示，这种双指数层次先验的形式与正态先验相比，能够获得更好的最小收敛速率。同时，双指数层次先验的形式为不同的系数设定不同的方差，能够根据数据特点自适应调整数据的离散程度，在后续的实验中对此有讨论。

对于参数 $\boldsymbol{\omega}$，常见的先验为 $p(\boldsymbol{\omega}) = q^{|\boldsymbol{\omega}|}(1-q)^{K-|\boldsymbol{\omega}|}$，其中 q 是超参数。这种先验的前提是假设 $\boldsymbol{X}$ 中的列向量之间是不相关的，每个列向量以 q 概率对观测值 $\boldsymbol{y}$ 不产生影响，或以概率 $qf(\boldsymbol{\beta})$ 对观测值产生影响。根据式(3.34)，$\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\omega},\sigma^2,\tau$ 的联合分布为：

$$p(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{\omega},\sigma^2,\tau) \propto \exp\left(\frac{\sum_{i=1}^{N} -(y_i - X\beta_i)^2 + 2\sigma^2\tau\sum_i|\beta_i|}{2\sigma^2}\right) q^{|\boldsymbol{\omega}|}(1-q)^{K-|\boldsymbol{\omega}|} \tag{3.35}$$

如果通过最大后验的方式计算 ω，需要对所有 β_i 进行积分，再通过二类最大似然估计(Type II MLE)获得 $\hat{\omega}$，使其满足 $\underset{\omega}{\operatorname{argmax}} \log p(\omega \mid \boldsymbol{y})$，但此过程中高维积分的计算不那么容易。

我们采用贝努利-贝塔过程对 ω 进行先验设置，ω_k 服从贝努利过程：

$$\begin{aligned} \omega_k \mid B &\sim \text{BeP}(B),\ k = 1,2,\cdots \\ B \mid \alpha, B_0 &\sim \text{BP}(\alpha, B_0) \end{aligned} \tag{3.36}$$

如上文所述，假设 $\boldsymbol{X}$ 中列向量之间不相关，则 β_k 的选取是相互独立的，亦即 ω_k 之间也是独立的。考虑到 $\boldsymbol{\beta}$ 稀疏度的不确定性和有界性，采用离散的贝塔过程。

回顾第 2 章中描述的离散贝塔过程，其产生的原子点与基础测度 B_0 的原子点位置相同，权重是以基础测度权重为参数的贝塔分布。我们以 Stick-breaking 过程构建离散的贝塔过程，具体过程如下：

$$\begin{aligned}
&\omega_k \sim \text{Bernoulli}(\theta_k) \\
&\theta_k = \sum_i^{c_i} p_i \delta_\omega \\
&c_k \sim \text{Poisson}(\frac{1}{\nu}k) \\
&p_i \sim \text{Beta}(\alpha q_i, \alpha(1-q_i)) \\
&q_i \sim V_i \prod_{j=1}^{i-1}(1-V_i)\delta_{\omega_i} \\
&V_i \sim \text{Beta}(1,\alpha)
\end{aligned} \tag{3.37}$$

在构建过程中，q_i 由 Stick-breaking 过程获得，从而使得 $q_i \leqslant 1$，$\sum_i q_i \leqslant 1$，这满足离散贝塔过程中对基础测度的需要，$B_0 = \sum_i q_i \delta_{\omega_i}$；采样的原子个数由带有参数 $\frac{1}{\nu}k$ 的泊松分布获得，从而使每轮原子的采样在有限的空间内获得；对于构建过程的收敛性，$B = \sum_i p_i \delta_{\omega_i}$，其中 $p_i \sim \text{Beta}(\alpha q_i, \alpha(1-q_i))$，能够以 Lévy 测度对其描述：

$$L(\mathrm{d}p, \mathrm{d}\omega) = \sum_i \text{Beta}(\alpha q_i, \alpha(1-q_i)(\mathrm{d}p)\delta_{\omega_i}(\mathrm{d}\omega) \tag{3.38}$$

构造过程符合 Lévy 过程。根据 Lévy 过程收敛定理，构造过程是收敛的。

对于集中参数 α 的选取，考虑到提高 $\boldsymbol{\beta}$ 的稀疏度，希望贝努利分布中 θ_k 尽可能为 0，鉴于贝塔分布 $\text{Beta}(a,b)$ 的均值为 $\frac{a}{a+b}$，选择 $\alpha > 1$。同时，在 $\alpha > 1$ 时，V_i 的值趋向于 0，使得 q_i 的值更靠近 0，这更进一步使得 p_i 的均值趋向于 0。

3.3.2 推理过程

本章采用 Gibbs 采样的方法获得后验：

$$(\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\theta}, \sigma^2, \tau \mid \boldsymbol{y}) \tag{3.39}$$

以 Gibbs 采样的方法获得的这个序列通常会快速收敛到 $\boldsymbol{\omega} \sim p(\boldsymbol{\omega} \mid \boldsymbol{y})$，不需要进行整体的后验计算，而且，更重要的是，这个序列中包含所需要的解，因为概率高的 ω_i 一定经常出现，而不经常出现的分量则可以丢弃。

通过 Gibb 采样的方法生成如下的序列：

$$\beta^{(0)}, (\sigma^2)^{(0)}, \omega^{(0)}, \theta^{(0)}, (\tau^2)^{(0)}, \beta^{(1)}, (\sigma^2)^{(1)}, \omega^{(1)}, \theta^{(1)}, (\tau^2)^{(1)}, \cdots$$

具体采样过程如下：

(1)对 $\boldsymbol{\beta}$ 进行采样：

$$\begin{aligned}\beta^{(j)} &\sim f(\boldsymbol{\beta} \mid \boldsymbol{\omega},\boldsymbol{\theta},\sigma^2,\tau^2,\boldsymbol{y}) \\ &= N(\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{y},\sigma^2\boldsymbol{A}^{-1})\text{Bernoulli}(\boldsymbol{\omega};\boldsymbol{\theta})\end{aligned} \tag{3.40}$$

其中 $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{D}_\gamma^{-1}$，$\boldsymbol{D}_\gamma = \text{diag}(\gamma_1^2,\cdots,\gamma_N^2)$。

(2)对 σ^2 进行采样：

$$\begin{aligned}(\sigma^2)^{(j)} &\sim f((\sigma^2)^{(j)} \mid \boldsymbol{y},\boldsymbol{\beta}) \\ &= \text{IG}\left(\frac{N+a_0}{2},\frac{|\boldsymbol{y}-\boldsymbol{X\beta}|^2+b_0}{2}\right)\end{aligned} \tag{3.41}$$

(3)对 $\boldsymbol{\theta}$ 进行采样：

根据 Beta 过程

$$(\theta_k)^j = \text{Beta}\left(\alpha + \sum_{i=1}^{N}(\omega_{ik})^{(j-1)},\alpha + N - \sum_{i=1}^{N}(\omega_{ik})^{(j-1)}\right) \tag{3.42}$$

(4)对 $\boldsymbol{\omega}$ 进行采样：

$$(\omega_k)^j \sim f((\omega_k)^j \mid \boldsymbol{y},\beta^j,(\sigma^2)^{(j)},\omega_{-i}^j,\theta^j) = f((\omega_k)^j \mid \beta^j,(\sigma^2)^{(j)},\omega_{-i}^j,\theta^j) \tag{3.43}$$

其中 $\omega_{-i}^j = (\omega_1^j,\cdots,\omega_{i-1}^j,\omega_{i+1}^j,\cdots,\omega_K^j)$。这个分布与观测值 $\boldsymbol{y}$ 无关，因为 $\boldsymbol{\omega}$ 通过 $\boldsymbol{\beta}$ 对 $\boldsymbol{y}$ 产生影响。式(3.43)的分布是贝努利分布，概率分布函数如下：

$$P((\omega_k)^j = 1 \mid \beta^j,(\sigma^2)^{(j)},\omega_{-i}^j,\theta^j) = \frac{c}{c+d} \tag{3.44}$$

其中

$$\begin{aligned}c &= \theta_k^j f(\beta^j \mid \omega_{-i}^j,(\omega_k)^j = 1)\theta_k^j \\ d &= \theta_k^j f(\beta^j \mid \omega_{-i}^j,(\omega_k)^j = 0)(1-\theta_k^j)\end{aligned} \tag{3.45}$$

(5)对 γ_k^2 进行采样：

$$1/(\gamma_k^2)^j \sim \text{IG}\left(\sqrt{\tau^2(\sigma^2)^j/((\beta_k)^j)^2},\tau^2\right)$$

其中 $\tau^j = \sqrt{2K/\sum_{i=1}^{N}(\omega_i k)^j}$。

具体算法如表 3.2 所示。

表 3.2　DMPBP 算法

算法 1：DMPBP 算法
输入：$\boldsymbol{X}$、$\boldsymbol{y}$、α，a_0，b_0 输出：$\boldsymbol{\beta}$ (1)初始化：令所有的 $\gamma_i = 1$，$\omega_i = 0$，开始迭代； (2)第 j 次迭代，第 1 步：根据式(3.40)得到 采样； (3)第 2 步：根据式(3.41)得到 $(\sigma^2)^{(j)}$ 采样； (4)第 3 步：根据式(3.42)得到 θ 采样； (5)第 4 步：根据式(3.44)和 (3.40)，计算 ω； (6)第 5 步：根据式(3.46)得到 γ 采样； (7)迭代(2)～(6)直至收敛。

3.3.3 人工信号实验结果与分析

为验证模型的性能，首先通过稀疏信号重构中常用的单元脉冲信号的重构实验进行测试。采用单位脉冲信号进行重构实验，可以在可控的不同信号长度、稀疏度、噪声等条件下，比较各种方法的重构效果。

实验首先生成信号长度为 $N = 512$，其中包含 $M = 20$ 个峰值，峰值所在位置以均匀分布随机选择，峰值为 1。如图 3.5(a)所示，其中数值 1 表示脉冲信号强度。再为生成的脉冲信号加上噪声，以模拟实际信号传输中产生的噪声影响。在实验中，噪声信号符合 $N(0, 0.005^2)$，得到的观测信号如图 3.5(b)所示。实验的目标是能够根据观测信号重构原始脉冲信号的脉冲峰值，并给出重构的误差，重构信号的峰值越接近原始信号的峰值，且误差越小，则表明模型的效果更好。我们对本章提出的 DMPBP 模型和其他方法进行比较实验，实验中过完备矩阵由 SVD 分解获得，重构的错误率由 $\| \hat{w} - w \|_2^2 / \| w \|_2^2$ 进行计算，其中 $\hat{w}$ 和 w 分别表示估计向量和真实向量。

首先通过基追踪算法和正交匹配追踪算法对信号进行重构，如图 3.5(c)和图 3.5(d)所示，再通过基于 RVM、贝叶斯 Lasso 和 DMPBP 的信号重构，如图 3.5(e)、(g)所示，这三种方法均通过差错线来表示协方差的偏差。与基追踪算法和正交匹配追踪算法相比，三种基于贝叶斯方法的压缩感知算法均能够对单元脉冲信号进行较为一致的重构。此外，贝叶斯压缩感知还提供了对于未知信号求解后验概率的方法，而不是点的估计，这个概率分布估计可以用协方差矩阵对系数进行估计。

对一维脉冲信号进行重构实验，实验结果如图 3.6 所示。对观测信号 $\boldsymbol{y}$ 的维度 N 从左到右、从上到下分别设置为：100～190，300～420，500～590，800～100，

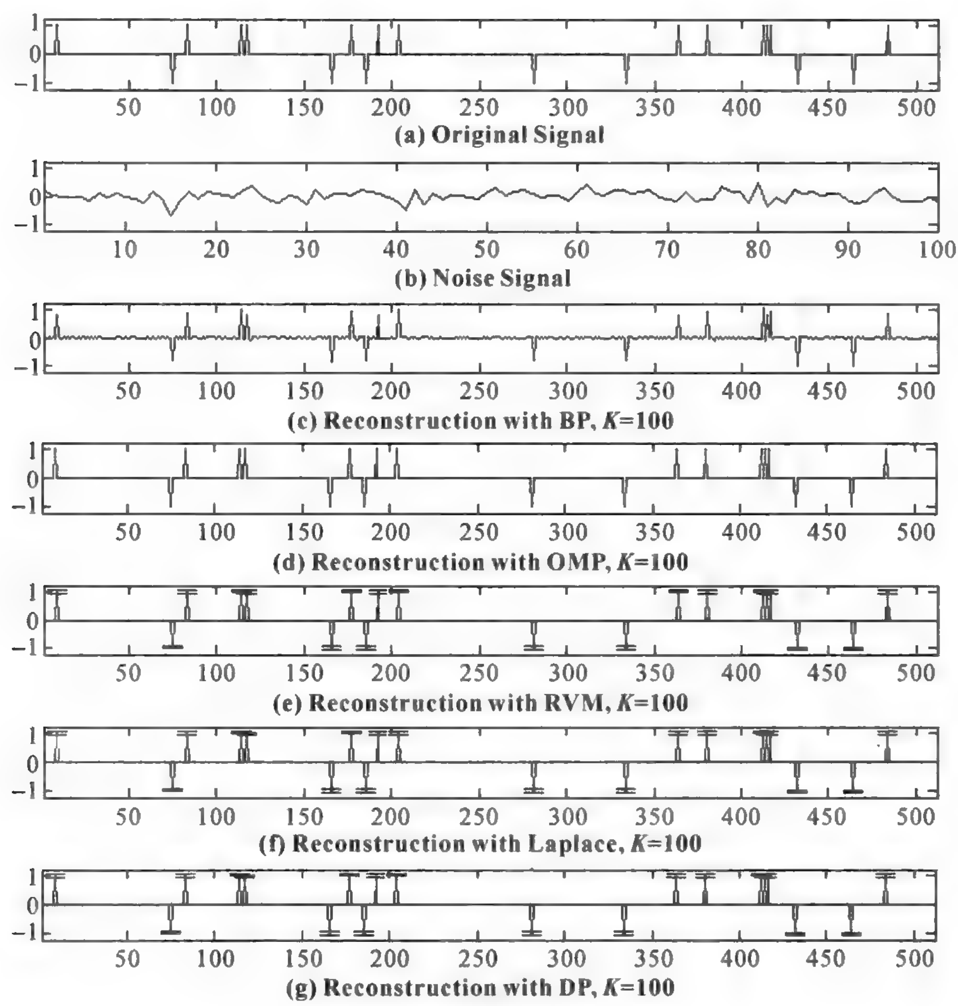


图 3.5　在单元脉冲信号上的重构结果

以观察不同维度区间段中各算法的重构误差。实验中假设原始信号稀疏度为 $N/2$。由实验结果可以看出，BCS 在不同维度下的误差都表现为最大，DMPBP 算法与基追踪算法的重构误差大多数小于 1，且两者之间的差距较小，尤其在观测信号维度较大的情况下两种算法的重构误差几乎相同。

接下来，在相同观测信号维度、原始信号稀疏度不同的条件下进行实验，设置稀疏度分别为 $N\times0.05$，$N\times0.2$，$N\times0.5$，$N\times0.8$ 进行比较，实验结果见图 3.7。对于稀疏度较低的情况，如图 3.7 中左上图所示，DBP 算法的效果与基追踪和 BCS 相比，重构误差较大。随着稀疏度的增加，DMPBP 算法的重构误差小于其他两种算法。

接下来的实验对一维正态分布随机信号进行重构，这更接近于多数问题的分布假设。实验中由 $N(0,8)$ 生成随机信号，通过稀疏度参数控制，设定其中有部分值为 0。图 3.8 显示了不同稀疏度下一维高斯随机信号的重构误差，稀疏度分别

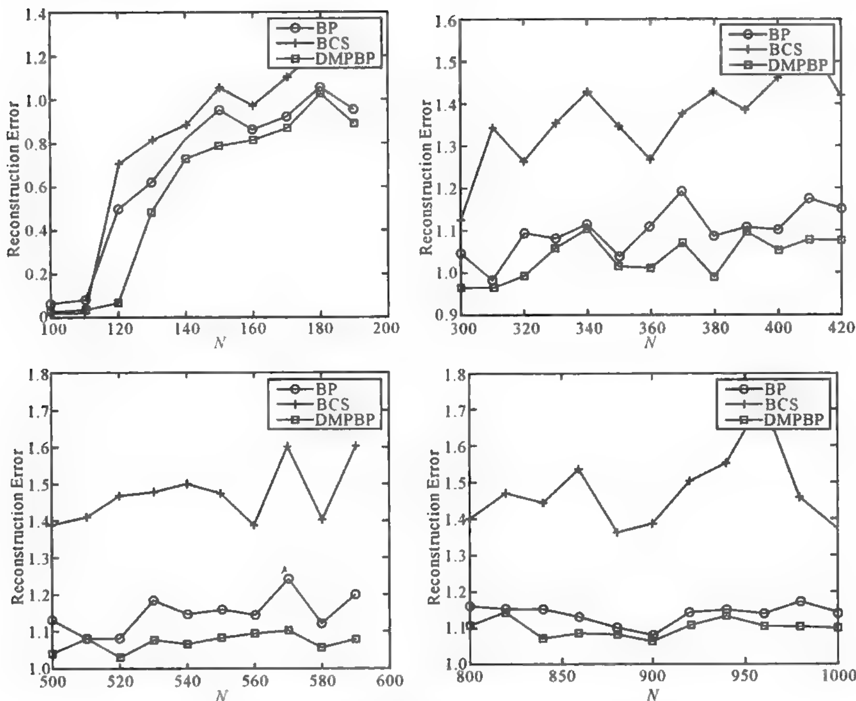


图 3.6 对不同维度的一维脉冲信号的重构误差比较

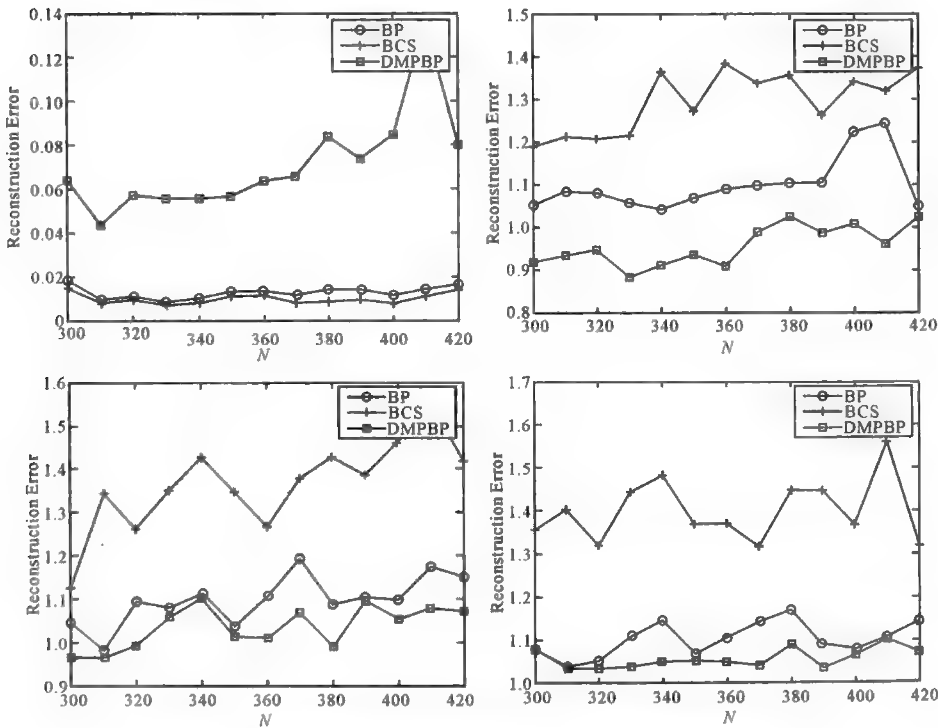


图 3.7 对稀疏度不同的一维脉冲信号重构误差比较

为 $N\times0.05$，$N\times0.2$，$N\times0.5$，$N\times0.8$。由图可以看出，在不同的稀疏度下，与基追踪和 BCS 相比，DMPBP 算法均取得更好的重构误差。

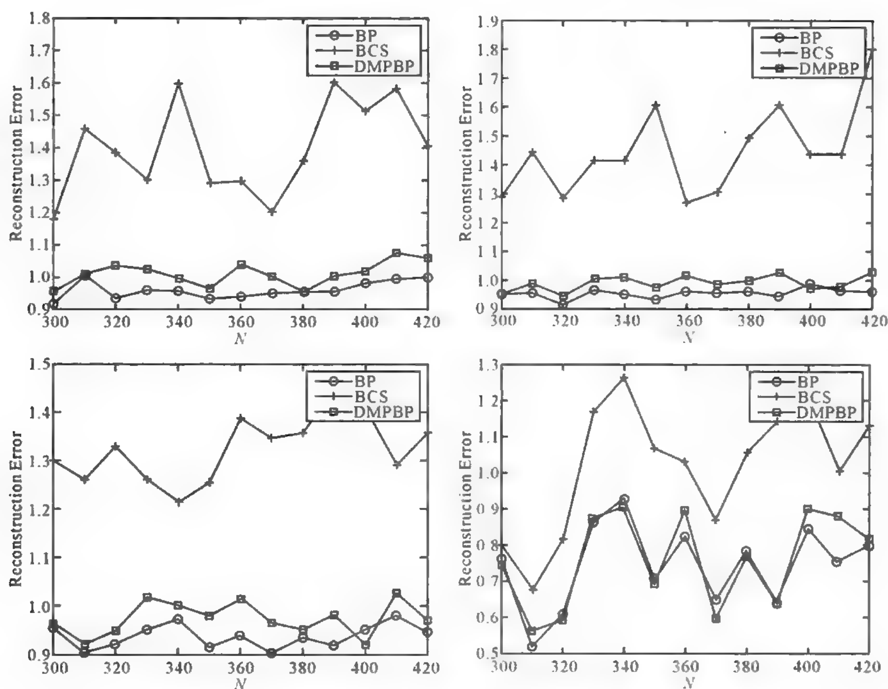


图 3.8 对稀疏度不同的一维高斯随机信号重构误差比较

为了验证 τ 的初值对模型的影响，分别选择 $\tau = 0, 1, 10$，对不同稀疏度的一维高斯随机信号进行重构实验，实验结果如图 3.9 所示。τ 初值为 0 得到的信号重构误差远大于其他值得到的重构误差，但 $\tau = 1$ 与 $\tau = 10$ 之间的信号重构误差相差较小。

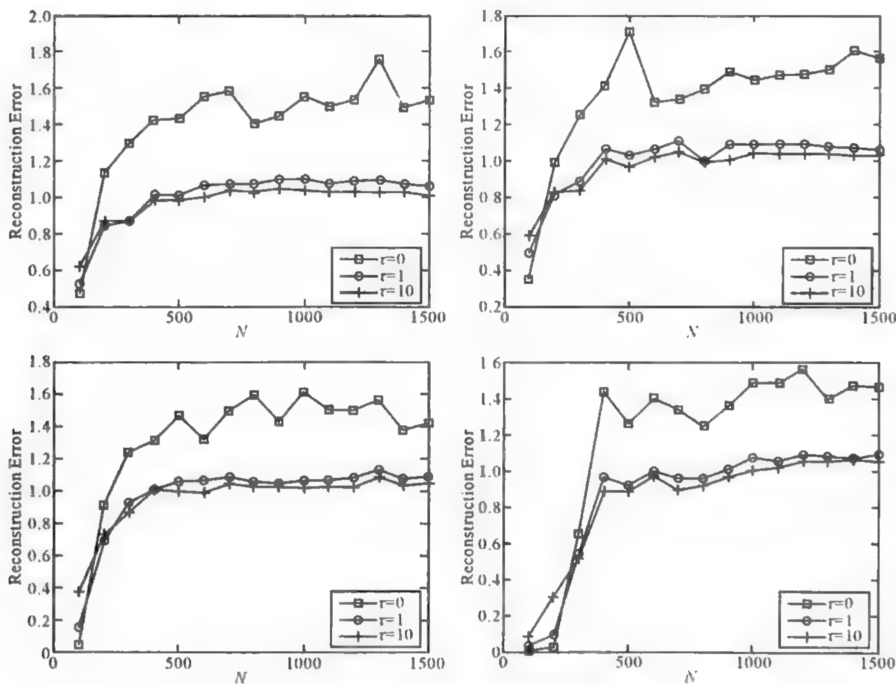


图 3.9 τ 的初值对信号重构误差的影响

3.3.4 手写数字识别实验结果

在上一节模拟信号重构分析的基础上，针对 DMPBP 模型对手写数字的识别进行实验。数据采用 USPS 美国邮政服务手写数字识别库，库中均为 16 × 16 像素的灰度图像的值，灰度值已被归一化。库中共有 9298 个手写数字图像，其中 7291 个用于训练，2007 个用于测试。训练时采用本书下一章的字典学习方法获得特征字典，每个数字对应的特征字典包含 64 种数字的形态，特征字典元素按照其对训练集的影响权重降序排列，如图 3.10 所示。

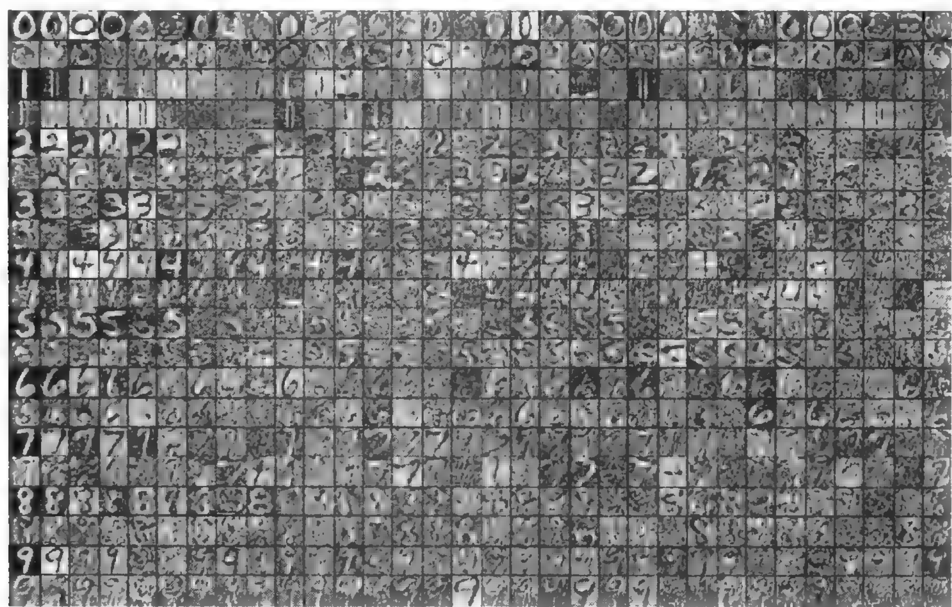


图 3.10　数字特征字典

实验主要针对数字“3”“5”进行识别，对每个测试集图像取其在特征字典上计算 100 次得到的平均稀疏表示概率。5 幅手写数字“3”“5”及其通过 DMPBP 算法计算所得的平均概率如图 3.11 和图 3.12 所示。从实验结果可以看出，算法对于数字 “3”“5”拐点特征明显的数字识别率较高。

3.4 小　结

本章通过扩展稀疏向量的函数形式，针对稀疏度根据观测数据自适应调整的需要，使用一种利用离散混合先验贝塔过程进行稀疏表示的方法。该方法能够根据观测数据在已知测量矩阵上的稀疏投影频率调整稀疏向量的稀疏度，并且模型中以高斯分布表示的拉普拉斯先验逼近 l_0 范数的方法，能够进一步提高稀疏表示

的能力和计算速度。

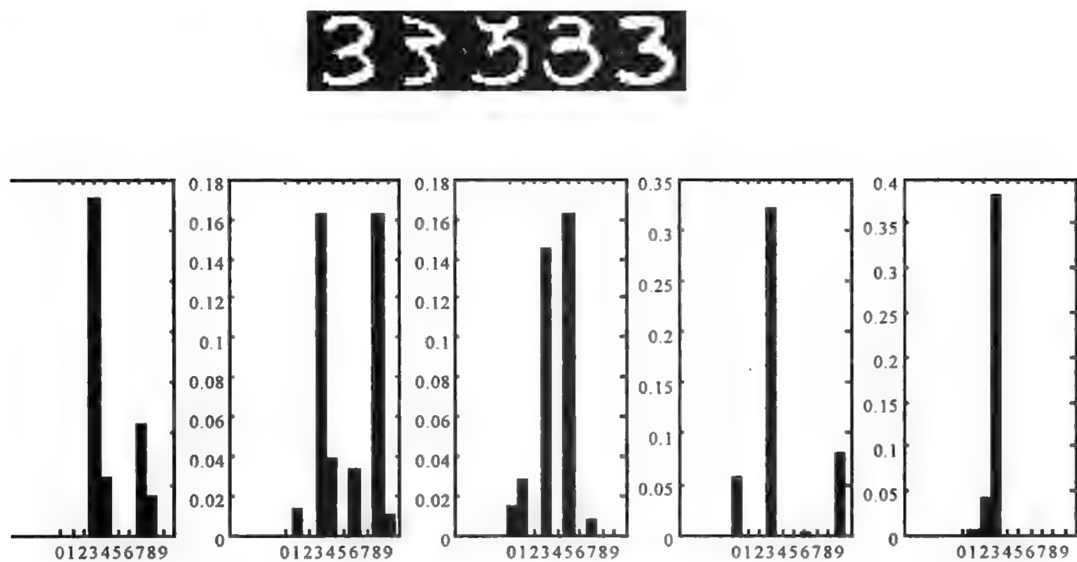


图 3.11　数字“3”的测试图像及其概率表示

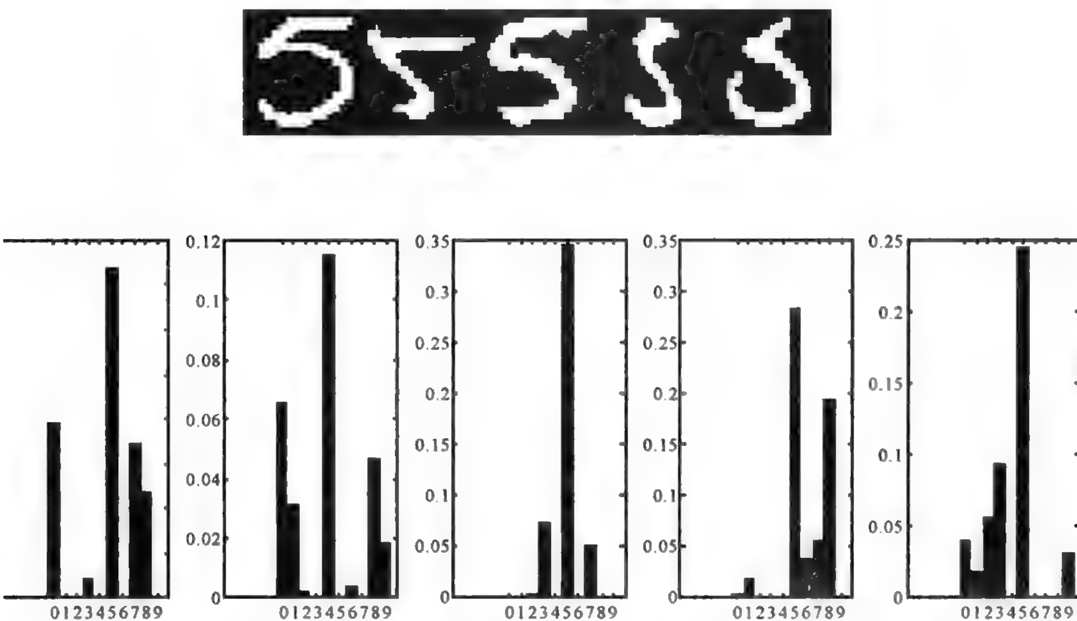


图 3.12　数字“5”的测试图像及其概率表示

第4章
基于聚类特征的贝叶斯非参数字典学习

在稀疏表示问题中，选择合适的基是保证信号稀疏度的基础和前提。1993年，Mallat等人提出过完备冗余字典对信号进行稀疏分解的思想，为稀疏表示的研究提供了新的解决思路。他们通过对自然语言的分析，说明过完备字典对信号表示的必要性，强调字典的构成应较好地符合信号本身所固有的特性，以实现匹配追踪算法的自适应分解。从另一方面，“稀疏”使得过完备字典成为可能。

基于过完备字典的信号稀疏表示问题在信号处理、压缩感知和特征提取等领域展示了让人印象深刻的效果。在最初的求解中，过完备字典多数为事先构建，例如通过Wavelets，Curvelets，SVD分解等方式，这些过完备字典在稀疏信号求解的过程中不发生变化。合理构建的过完备字典能够引发稀疏信号的快速求解，但这种结构固定的过完备字典不能够自适应观测数据的变化。逐渐地，稀疏表示问题中过完备字典的生成引起研究人员的关注，用学习的字典，而不是提前生成的字典（例如，小波字典）中的原子的线性组合表现信号，在低层图像处理任务（例如降噪）中取得了良好的效果。在Candès和Tao对压缩感知的描述和证明后，过完备字典的学习成为稀疏问题的研究热点。

本章对采用贝叶斯非参数方法构建字典进行分析和研究，在前一章稀疏表示的贝努利-贝塔过程建模的基础上，给出一种基于高斯过程聚类的贝叶斯非参数字典学习方法。实验结果证实了该方法的有效性，并与其他的字典学习方法相比，该方法在模型精度、稀疏度和字典维度的自适应性上有一定优势。

4.1 字典学习问题

字典学习问题的描述如下：

对于观测信号 $\boldsymbol{y} \in \mathbf{R}^N$，如果将其分解为

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{4.1}$$

其中 $\boldsymbol{\varepsilon}$ 是符合 $N(0,\sigma^2 I)$ 的噪声。式(4.1)中 $\boldsymbol{X} \in \mathrm{R}^{N\times K}$ 和 $\boldsymbol{\beta} \in \mathbf{R}^K$ 均未知，且要得到最稀疏的 $\boldsymbol{\beta}$，则有

$$\min_{X,\beta} \| \boldsymbol{\beta} \|_0 \quad \text{s.t.} \ \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{4.2}$$

式(4.2)描述的问题被称为字典学习(Dictionary Learning)。

在字典学习问题中，基函数被超完备的冗余函数库所取代，这个冗余函数库通常被称为冗余字典，简称“字典”。字典中的元素被称为原子。字典的选择应尽可能好地符合被逼近信号的结构，字典的构成可以没有任何限制。从冗余字典中找到具有最佳线性组合的 K 项原子来表示一个信号，称为信号的稀疏逼近或高度非线性逼近。

简而言之，字典学习问题包含两个方面：①构造尽可能好的过完备字典 $\boldsymbol{X}$。②从这个过完备字典中拣选最好的若干项的组合来描述观测信号，给出信号基于 $\boldsymbol{X}$ 的稀疏表示。

字典学习涉及很多方面的知识与应用。神经科学的理论研究指出，基于过完备字典的稀疏表示更符合哺乳动物视觉系统的生物学背景。非线性逼近理论也从理论上证明了基于过完备字典对信号的逼近要优于正交基。在图像处理方面，基于过完备字典能获得图像的稀疏表示，已应用于图像处理的各个领域。

字典学习研究的问题主要包括以下几个方面：字典学习理论的研究、基于过完备字典的稀疏表示快速算法的研究、过完备字典的构造研究及应用领域的研究等。

4.2 现有字典学习算法

在字典学习问题被提出后，字典学习算法层出不穷。字典学习包含的两个方面内容为字典学习提供了一种直观的方法，即采用交替更新字典 D 和稀疏向量 β 的方法，得到字典和稀疏信号的收敛值。

主要方法有贪婪法、最大似然或最大后验法、在线字典学习法等，这些方法都采用字典和稀疏向量交替优化的方法进行求解。

4.2.1 贪婪法

1)最佳方向方法

最佳方向方法(Method of Optimal Directions，MOD)在 GLA(Generalized Lloyd Algorithm)算法基础上对向量和原子进行优化计算。算法在初始化时从输入信号中随机地选取 K 个列向量，并对字典原子进行规范化处理；在迭代过程中，

根据当前字典，利用 OMP 算法更新稀疏向量；在更新字典时，采用最小二乘法根据当前稀疏向量更新字典，使式(4.3)最小。

$$\|\boldsymbol{\varepsilon}\|_F^2 = \|\boldsymbol{y} - \boldsymbol{X\beta}\|_F^2 \tag{4.3}$$

其中 $\|\cdot\|_F^2$ 表示 $\|\boldsymbol{A}\|_F^2 = \sqrt{\sum_{ij} A_{ij}^2}$，$\boldsymbol{\beta} = [\beta_1, \cdots \beta_P]$，得到 $\boldsymbol{X}$ 的更新公式为：

$$\boldsymbol{X}^{i+1} = \boldsymbol{y\beta}^{(i)^T} \cdot (\boldsymbol{\beta}^{(i)} \boldsymbol{\beta}^{(i)^T})^{-1} \tag{4.4}$$

在更新字典过程中，如果字典中某原子二范数接近于0，为了进行下一轮迭代计算应该忽略该列原子，则重新从输入信号中随机地选取1个列向量代替该原子，重复迭代操作直到收敛。

2）K-SVD 算法

K-SVD 算法是目前字典学习中最为流行的算法之一。在求解稀疏向量时，同 MOD 一样，根据当前所得的过完备字典 D，采用 OMP 算法计算信号在字典上的稀疏系数。K-SVD 算法的字典与 MOD 不同。K-SVD 算法在固定 B 求优化字典的过程中，对字典中的原子 x_k 依次更新。从当前获得的 B 中，取与字典第 k 列原子相应的 k 行构成第 β_T^k 行向量，通过

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{X\beta}\|_F^2 &= \|\boldsymbol{y} - \sum_{j=1}^{K} x_j \beta_T^j\|_F^2 \\ &= \|\boldsymbol{y} - \sum_{j \neq k} x_j \beta_T^j - x_k \beta_T^k\|_F^2 \\ &= \|\boldsymbol{\varepsilon} - x_k \beta_T^k\|_F^2\end{aligned} \tag{4.5}$$

把 $\boldsymbol{X\beta}$ 降为 $K-1$ 的矩阵，固定 $K-1$ 个对象求第 k 个字典原子和相应的 β_T^k。取 β_T^k 中不为0的那些分量构成向量 β_R^k，β_R^k 的维度小于等于 β_T^k。相应地得到与 β_R^k 对应的 y_k^R 和 $\boldsymbol{\varepsilon}_k^R$，则式(4.5)可表示为

$$\|\boldsymbol{\varepsilon}_k^R - x_k \beta_R^k\|_F^2 \tag{4.6}$$

将式(4.6)中 $\boldsymbol{\varepsilon}_k^R$ 用 SVD 将其分解为 $\boldsymbol{\varepsilon}_k^R = U\Delta V^T$，得到的 U 中的第一列为 $\tilde{x}_k$，V 中的第一列为 β_R^k。字典的一次更新需要进行 K 次 SVD 分解。正是因为如此，算法得名为 K-SVD。

K-SVD 每次更新其字典原子和其对应的稀疏系数，直到所有的原子更新完毕。重复迭代直到收敛就得到优化的字典和稀疏系数。

采用贪婪法通常可以得到最优解，但由于迭代次数和收敛速度的影响，算法运行速度较慢。研究者对 K-SVD 算法的改进和优化提出了不少新的方法，从一定程度上提高了算法的计算速度。

4.2.2 贝叶斯方法

1)最大似然估计法

基于最大似然估计的字典学习算法采用梯度优化(Gradient Optimization)的方法,在对字典的更新过程中,通过对原子进行规范化处理来降低误差。

对于 T 个相互独立的观测信号 $\boldsymbol{y}^T=(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_T)$,字典 X 的最大似然估计模型为

$$\hat{\boldsymbol{X}}_{ML}=\underset{\boldsymbol{X}}{\operatorname{argmax}}\, p(\boldsymbol{Y};\boldsymbol{X})=\underset{\boldsymbol{X}}{\operatorname{argmax}}\prod_{t=1}^{T}p(\boldsymbol{y}_t;\boldsymbol{X}) \tag{4.7}$$

其中

$$p(\boldsymbol{y}_t;\boldsymbol{X})=\int p(\boldsymbol{y}_t,\boldsymbol{\beta};\boldsymbol{X})\mathrm{d}\boldsymbol{\beta}=\int p(\boldsymbol{y}_t\mid\boldsymbol{\beta};\boldsymbol{X})p(\boldsymbol{\beta})\mathrm{d}\boldsymbol{\beta} \tag{4.8}$$

式(4.7)也可以写为:

$$\hat{\boldsymbol{X}}_{ML}=\underset{\boldsymbol{X}}{\operatorname{argmin}}-\sum_{t=1}^{T}p(\boldsymbol{y}_t;\boldsymbol{X}) \tag{4.9}$$

对此模型的计算,需要根据对 β 和 ε 的假设先验对 T 个未知稀疏向量 β 进行积分,但这个积分运算通常是不易处理的或计算上是不可行的,因此,需要对积分采用逐步逼近的计算方法。

2) 最大后验估计(MAP)

基于最大后验估计的字典学习算法采用贝叶斯模型,在交替更新过程中计算稀疏向量和字典原子的最大后验来获得最优解。

未知信号 $\boldsymbol{\beta}$ 的似然函数为:

$$\begin{aligned} p(\boldsymbol{\beta}) &\approx \delta(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{X}})),\\ \hat{\beta}(\hat{\boldsymbol{X}}) &= \underset{\hat{\beta}}{\operatorname{argmax}}\, p(\boldsymbol{y}_t,\hat{\boldsymbol{\beta}};\boldsymbol{X}) \end{aligned} \tag{4.10}$$

其中 $\hat{\boldsymbol{X}}$ 是对 $\boldsymbol{X}$ 的当前估计值。在此假设下,式(4.9)变为

$$\hat{\boldsymbol{X}}_{ML}=\operatorname{argmin}-\sum_{t=1}^{T}\|\boldsymbol{y}_t-\hat{\boldsymbol{X}}\hat{\boldsymbol{\beta}}\|^2/2\sigma^2 \tag{4.11}$$

Kenneth 等人利用式(4.10)的逼近方法,在其提出的求解稀疏表示的算法 FOCUSS 基础上,给出了不同约束条件下字典 $\boldsymbol{X}$ 的最大后验估计的近似逼近算法。Yaghoobi 等人提出一种对字典采用了凸松弛约束的方法,采用了 Majorization Minimization Algorithm 对字典和稀疏向量进行交替更新。

4.2.3 在线方法

Mairal 等人根据凸优化理论,对由 T 个采样 $\boldsymbol{y}_t$ 构成采样矩阵同时求解稀疏矩阵 β_t 和最优字典 $\boldsymbol{X}$ 进行分析,证实最优字典和稀疏向量是非凸的,而将问题分为固定字典求稀疏向量、固定稀疏向量求最优字典则分别是凸的。因此,采用交替求解稀疏因子和字典的方法,利用矩阵分解对字典学习问题进行求解。该文对字典

的约束为：

$$\{\boldsymbol{X} \in \boldsymbol{C}, \mathrm{C} \triangleq \boldsymbol{X} \in \mathbf{R}^{N \times K} \quad \text{s.t.} \ \ \forall j = 1, \cdots, K, \quad d_j^T d_j \leqslant 1\} \tag{4.12}$$

在此约束下利用一阶投影随机梯度下降算法对字典进行更新

$$\boldsymbol{X}_t = \Pi_C[\boldsymbol{X}_{t-1} - \delta_t \nabla_X l(y_t, \boldsymbol{X}_{t-1})] \tag{4.13}$$

并对字典的收敛性进行了证明，并通过图像降噪实验展示了该算法在获得同样降噪效果的条件下，计算速度优于 K-SVD。

4.2.4 非参数方法

上述方法均假设字典矩阵的维度 K 固定不变，如果 K 随着观测数据发生变化，需要借助非参数方法。Zhou 等人给出一种基于贝塔过程的非参数方法 BPFA，为字典矩阵中各元素设置均值为 0，方差为 $1/N$ 的先验，通过 Gibbs 采样计算字典原子的后验。尽管该方法没有从理论上进行证明其生成的冗余矩阵满足 RIP 条件，但图像降噪实验表明高斯先验能够反映图像的数据特征，与 DCT，K-SVD 算法相比，基于 Beta 过程的字典学习算法对图像降噪的效果优于 DCT 字典，与 K-SVD 算法效果相当。

4.3 约束等距性条件

2001 年 Donoho 等人对 l_0 范数和 l_1 范数基于两个标准正交基构成的联合字典具有相同唯一解进行讨论，即对于两个标准正交基 Φ, Ψ 构成的字典 $[\Phi, \Psi]$，如果观测信号 y 可以由稀疏信号 $\boldsymbol{x}$ 表示，即 $\boldsymbol{y} = [\Phi, \Psi]\boldsymbol{x}$，且 $|\boldsymbol{x}|_0 < 0.9142/M$，则 l_1 范数最小解和 l_0 范数最小解一致，其中 $M = \sup\limits_{1 \leqslant i,j \leqslant N}(|\phi_i, \psi_j|)$。

Candès 和 Tao 证明了过完备字典必须满足约束等距性条件（Restricted Isometry Property，RIP)，即对于任意 $\boldsymbol{c} \in \mathbf{R}^{|S|}$ 和常数 $\delta_S \in (0,1)$，如果

$$(1 - \delta_S)\|\boldsymbol{c}\|_2^2 \leqslant \|\boldsymbol{c}\widetilde{\boldsymbol{X}}_S\|_2^2 \leqslant (1 + \delta_S)\|\boldsymbol{c}\|_2^2 \tag{4.14}$$

成立，其中 $T \subset \{1, \cdots, K\}$，$\|T\| \leqslant S$，$\widetilde{\boldsymbol{X}}_S$ 为 $\boldsymbol{X}$ 中由索引 S 所指示的相关列构成的大小为 $N \times \|S\|$ 的子矩阵，则称矩阵 $\boldsymbol{X}$ 满足约束等距性。通常，对于一个 S 维稀疏信号 $\boldsymbol{\beta}$，可以从观测信号 $\boldsymbol{y}$ 精确重构 $\boldsymbol{x}$ 的充分条件是矩阵 $\boldsymbol{X}$ 对于任意 $\boldsymbol{c} \in \mathbf{R}^{|S|}$ 和常数 $\delta_{2S} \in (0,1)$ 有 $2S$ 阶约束等距性，即

$$(1 - \delta_{2S})\|\boldsymbol{c}\|_2^2 \leqslant \|\boldsymbol{c}\widetilde{\boldsymbol{X}}_S\|_2^2 \leqslant (1 + \delta_{2S})\|\boldsymbol{c}\|_2^2 \tag{4.15}$$

成立，其中 $T \subset \{1, \cdots, K\}$，$\|T\| \leqslant 2S$。

约束等距性条件表明，只要矩阵 $\boldsymbol{X}$ 满足该条件，那么将 K 维信号中最大的 S 个值稳定重建所需的采样为 $S \times \log(K/S)$。然而，尽管约束等距性条件拥有完美

的特性，但不能保证满足该条件的矩阵 $\boldsymbol{X}$ 存在。寻找满足约束等距性条件的矩阵 $\boldsymbol{X}$ 是统计稀疏学习和压缩感知问题中的重要研究内容。

4.4 带有聚类特征的贝叶斯非参数字典学习

通过现有算法对图像的字典学习的结果可以得知，图像中各原子 y_i 在经过字典 $\boldsymbol{X}$ 得到的稀疏表示向量 β_i 之间不是随机分布的，它们的位置不确定性通常与图像信号的非本地自相似相关，这就意味着，利用这种位置相关约束得到更为稀疏的表达的概率更高。聚类通常在处理这种非线性约束（位置相关）问题中取得突出的效果。其中董伟生等人为稀疏表示增加稀疏表示向量的 l_2 范数约束，并在计算中以 l_1 范数替代 l_2 范数以获得更优解。贝叶斯非参数方法对于聚类问题以其能够自适应得到类别特征获得极大的关注，然而在贝叶斯非参数中将稀疏表示和聚类共同构建相互促进的协同模型并不容易，一方面，稀疏表示和聚类分属不同运算等级的问题，另外，对稀疏表示和聚类构成的模型的后验计算并不容易。本章通过对图像降噪数据进行分析，给出一种带有聚类特征的贝叶斯非参数字典学习方法。

4.4.1 模型描述

近年来的研究表明通过一定概率分布独立同分布生成的（合理尺度的）随机矩阵能够以高概率满足约束等距性条件。Candès 在 2006 年证明了当矩阵 $\tilde{\Phi}$ 是高斯随机矩阵时，矩阵 $\boldsymbol{X}=\tilde{\Phi}\Psi$ 能够以较大概率满足约束等距性条件，其中 Ψ 是正交变换基。因此，我们通过选择一个大小为 $N\times\infty$ 的高斯矩阵得到字典 $\boldsymbol{X}$，字典中的值满足 $N(0,1/K)$ 的独立正态分布。

接下来，为字典进行先验约束。单位 Frobenius 范数对字典中所有原子进行约束，

$$\|\boldsymbol{X}\|_F^2=\mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X})=1 \tag{4.16}$$

尽管这种约束通过数学证明是可以得到稀疏解的，但约束过分松弛。因为采用单位 Frobenius 范数约束可能引发 $\boldsymbol{X}$ 中多项原子趋近于 0，原子的模很小，使得在求解稀疏表示中与该原子对应的稀疏系数很大，这些原子在稀疏分解中被充分利用而不断得到更新，而其他原子一直得不到更新。

与单位 Frobenius 范数约束相比，列规范化约束更严格。对字典 $\boldsymbol{X}$ 中每个原子 x_k：

$$\|\boldsymbol{x}_k\|^2=C,\ k=1,2,\cdots \tag{4.17}$$

即每个原子的模都是一个常量，使得在稀疏分解中为每个原子的选择赋予均匀

分布。

结合单位 Frobenius 范数约束和列规范化约束，设置字典先验为：

$$\| \boldsymbol{x}_k \|^2 = \frac{1}{K},\ k = 1,2,\cdots,K,\quad K \to \infty \tag{4.18}$$

对于字典维度 $K \to \infty$，贝叶斯非参数方法中估计的字典维度随着观测数据发生变化，因此，在对字典原子进行更新时，根据当前的字典维度 K 进行规范化约束。，

对稀疏表示的求解采用本文前一章的方法，即以贝努利-贝塔过程描述稀疏向量中的 0 分量以提高稀疏度，对 1 分量部分仍以层次高斯分布获得 Laplace 分布；对于 0 分量的部分，根据对图像数据的分析，采用均值为 0，以方差尽可能小的高斯分布来逼近。

在上述分析基础上，建立模型如下：

$$\begin{aligned}
&\boldsymbol{y}_i \sim N(\boldsymbol{X}\beta_i, \sigma^2 \boldsymbol{I}_N) \\
&\boldsymbol{x}_k \sim N(0, \frac{\alpha_0}{N}\boldsymbol{I}_N) \\
&\beta_i \sim (1-\omega_i)\delta_0(\beta_i) + \omega_i N(\mu_s, \lambda^{-1}\sigma^2 \boldsymbol{I}_K) \\
&\mu_s \sim \mathrm{GMM}(G_{c_n})
\end{aligned} \tag{4.19}$$

假设 $\delta_0(\beta_i)$ 用均值为 0，方差为 $\alpha^{-1}\sigma^2 \boldsymbol{I}_K$ 的高斯分布近似，其中 $\alpha_i^{-1} \sim \mathrm{IG}(1,\gamma/2)$，则将对式(4.19)中参数积分后得到

$$p(\boldsymbol{y}_i; \boldsymbol{X}, \beta_i) \propto \frac{1}{Z}\exp\{-\frac{1}{2\sigma^2}(\| \boldsymbol{y}_i - \boldsymbol{X}\beta_i \|^2 + \frac{N\sigma^2}{\alpha_0}\| \boldsymbol{x}_k \|^2 + \alpha_i^{-1}\ | \beta_i |\ + \lambda^{-1}\| \beta_i - \mu_{C_i} \|^2)\} \tag{4.20}$$

其中 Z 是标准化常量。似然函数中字典满足式(4.12)，稀疏向量 β_i 的表示既包含 l_1 范数约束，也包含其所属类的 l_2 范数约束。如果不考虑聚类约束，则 μ_s 均为 0，似然函数中对 β_i 的表示与弹性网对稀疏向量的约束类似。

式(4.19)中 ω_i 以 Bernoilli-Beta 过程建模：

$$\begin{aligned}
&\omega_i \sim \prod_{k=1}^{K} \mathrm{Bernoulli}(\theta_{c_n k}) \\
&\theta_{c_n k} \sim \mathrm{Beta}(a/K, b(K-1)/K)
\end{aligned} \tag{4.21}$$

对于模型中稀疏向量非 0 的分量，根据图像数据局部聚类的特点，采用高斯混合模型表示，但其中高斯混合模型的混合度是未知的，因此采用狄利克雷过程表示。式(4.19)中 c_n 作为类别的表示，采用多项式-高斯过程：

$$\begin{aligned}
&\boldsymbol{c}_n \sim \mathrm{Multinomial}(1,\cdots,S,\eta) \\
&G_s \sim \mathrm{DP}(cG_0) \\
&\eta \sim \mathrm{Dirichlet}(\frac{1}{S},\cdots,\frac{1}{S})
\end{aligned} \tag{4.22}$$

4.4.2 模型推理

对式(4.19)的求解,我们采用最大后验概率估计(MAP)和 Gibbs 采样算法(block Gibbs sampler)。在推理过程中,对于由贝努利-贝塔过程生成的 ω_i 和由多项式-高斯过程生成的 c_n,通过区域 Gibbs 采样的方法采样生成。对于模型中 x_i 和 β_i,则可以通过最大后验概率估计进行计算。在计算中,对于参数 σ^2,为其假设先验为 $\sigma^2 \sim \mathrm{IG}(a_0,b_0)$,其中超参数 a_0,b_0 之间的比值越大,则 σ^2 越接近于 0。对于参数 α_0,由于对字典的约束为标准列向量约束,$\boldsymbol{x}_i^T\boldsymbol{x}_i \leqslant 1$,因此,其对取值 $\alpha_0 \ll N$。λ 的选取与对稀疏向量的稀疏程度相关,模型中假设 λ 的先验分布为 $\mathrm{IG}(c_0,d_0)$,在采样中 λ 值不断根据观测数据进行更新。超参数 γ 作为 α^{-1} 的参数,α^{-1} 的期望接近于 0,所以 $\gamma \ll 2$。对于 Beta 过程中的超参数 a,b,其选取与稀疏向量的稀疏度有关,为获得尽可能稀疏的表示,a,b 的初值应满足 $\frac{a}{b} < K$,尽管 $k \to \infty$,在初始化过程中可以根据具体问题设定一个足够大的数 L,令 $\frac{a}{b} < L$。

为了计算的方便,令 P 个 $\boldsymbol{y}_i$ 构成 $N \times P$ 的矩阵 $\boldsymbol{Y}$。具体计算过程如下:

(1)更新 β_i:

对于 β_i,有:

$$p(\beta_i \mid \sim) \propto N(\boldsymbol{y}_i;\boldsymbol{X}\beta_i,\sigma^2\boldsymbol{I}_N)\prod_{i=1}^{K}\mathrm{Bernoulli}(\omega_{ik};\theta_k)$$
$$\{(1-\omega_{ik})N(0,\alpha^{-1}\sigma^2)+\omega_{ik}N(\mu_s,\lambda^{-1}\sigma^2)\} \tag{4.23}$$

通过最大后验概率估计更新 β_i 得到:

$$\beta_i = \{(\alpha+\lambda)\boldsymbol{I}_{KK} + \boldsymbol{X}^T\boldsymbol{X} \circ \omega_i\omega_i^T\}^{-1}\mathrm{diag}(\omega_i)(\boldsymbol{X}^T\boldsymbol{y}_i - \lambda\mu_s\boldsymbol{I}_k) \tag{4.24}$$

(2)更新 $\boldsymbol{X}$:

根据式(4.19)有:

$$p(\boldsymbol{x}_k \mid \sim) \propto \prod_{i=1}^{N}N(y_i;\boldsymbol{X}\beta_i,\sigma^2\boldsymbol{I}_N)N(x_k;0,\alpha_0N^{-1}\boldsymbol{I}_N) \tag{4.25}$$

根据最大后验概率估计更新 $\boldsymbol{X}$ 得到:

$$\boldsymbol{X} = \boldsymbol{y}_i\beta_i^T(\beta_i\beta_i^T + N\sigma^2/\alpha_0\boldsymbol{I}_K) \tag{4.26}$$

(3)更新 δ_{sk}:

根据贝塔过程的后验为式(2.46),得到 $\hat{\delta}_{sk}$:

$$\hat{\delta}_{sk} = \frac{\dfrac{a}{K} + \sum_{i=1}^{N}\omega_{is}\boldsymbol{I}(c_i = s)}{\dfrac{a}{K} + \dfrac{b(K-1)}{K} + \sum_{i=1}^{N}\boldsymbol{I}(c_n = s)} \tag{4.27}$$

(4)更新 η_s：

根据狄利克雷过程后验为(2.24)，得到 $\hat{\eta}_s$：

$$\hat{\eta}_s = \frac{\frac{1}{S} + \sum_{j=1}^{N} \boldsymbol{I}(c_n = s)}{N+1} \tag{4.28}$$

(5)更新 G_s：

G_s 的混合权重、均值和方差通过 $c_n = s$ 的β_i 计算，参考 Bishop 编写的 *Pattern Recognition and Machine Learning* 书中第 9 章中关于混合高斯的计算，得到：

$$\begin{gathered} N_s = \sum_{i=1}^{N} \boldsymbol{I}(c_n = s) \\ \mu_s = \frac{1}{N_s} \sum_{i=1}^{N} \omega_{is} \boldsymbol{I}(c_i = s) \boldsymbol{x}_i \\ \sigma_s = \frac{1}{N_s} \sum_{i=1}^{N} \omega_{is} \boldsymbol{I}(c_i = s)(\boldsymbol{x}_i - \mu_s)(\boldsymbol{x}_i - \mu_s)^T \\ \pi_s = \frac{N_s}{N} \end{gathered} \tag{4.29}$$

(6)更新 ω_i：

ω_i 由贝努利分布采样生成，

$$\omega_{ik} \sim \text{Bernoulli}\left(\frac{p_1}{p_0 + p_1}\right) \tag{4.30}$$

其中 p_1 是 $\omega_{ik} = 1$ 的概率，且

$$p_1 = \theta_k \exp\{-\frac{1}{2\sigma^2}(\beta_{ik} \boldsymbol{x}_k^T \boldsymbol{x}_k - 2\beta_{ik} \boldsymbol{x}_k^T \tilde{\boldsymbol{x}}_k)\} \tag{4.31}$$

其中 $\tilde{\boldsymbol{x}}_k = \boldsymbol{y}_i - \boldsymbol{X}\beta_i + \boldsymbol{x}_k \beta_{ik}$。

p_0 是 $\omega_{ik} = 0$ 的概率，且

$$p_0 = 1 - \theta_k \tag{4.32}$$

(7)更新 c_i：

c_i 由 S 维的多项式分布采样获得，其后验：

$$p(c_i = s \mid \sim) \propto p(\omega_i \mid \hat{\theta}_s) p(\beta_i \mid G_d) p(c_i = d \mid \hat{\eta}) \tag{4.33}$$

其中 $p(\omega_i \mid \hat{\theta}_s) = \prod_{k=1}^{K} \hat{\theta}_s^{\omega_i} (1 - \hat{\theta}_s)^{(1-\omega_i)}$，$p(c_i = d \mid \hat{\eta}) = \hat{\eta}(s)$，$p(\beta_i \mid G_s)$ 是根据 GMM G_d 计算的似然函数。

在上述推理演绎的基础上，算法 CLBP 如表 4.1 所示。

表 4.1 CLBP 算法

算法 2：CLBP 算法
输入：$\{y_1, \ldots, y_P\}$ 输出： (1)初始化：通过 SVD 得到； (2)根据式(4.24)计算，并根据式(4.26)计算当前值； (3)对每个 s，根据式(4.27)计算 δ_s，并通过式(4.29)得到 G_s 采样； (4)分别根据式(4.30)和(4.33)对 ω_i 和 c_i 采样； (5)迭代(2)～(4)直至收敛。

4.4.3 实验结果与分析

本章采用对灰度图像的降噪来验证模型的有效性，实验运行环境为四核 i5 280GHz 处理器，8GB 内存，matlab 版本为 R2010b。首先对 200 × 200 的方形棋盘灰度图和圆形棋盘灰度图进行降噪实验，噪声标准差 $\sigma = 50$，与参数方法中流行的 K-SVD 算法和贝叶斯非参数方法中基于贝塔过程的 BPFA 算法进行比较，结果如图 4.1 和 4.2 所示。图中第一行左侧为原图，右侧为加噪图像；第二行是通过不同字典得到的降噪后的效果图；第三行是通过不同算法得到的字典，由左到右分别是通过小波字典、KSVD 字典、BPFA 字典和 CLBP 字典，字典元素按照升序排列。

在对方形棋盘的降噪中，基于贝叶斯非参数的方法 BPFA 和 CLBP 效果都优于 K-SVD 算法，对圆形棋盘的降噪效果不及 K-SVD 算法。两种贝叶斯非参数方法之间，本章提出的 CLBP 算法优于 BPFA。

再对标准的灰度图进行比较，选择 256 × 256 的 Lena 图进行实验，噪声标准差分别选择 $\sigma = 25$ 和 $\sigma = 50$，实验结果如图 4.3 和 4.4 所示。

由实验结果可以看出，基于贝叶斯非参数的方法 BPFA 和 CLBP 均优于基于参数的 K-SVD 算法，且 CLBP 算法的效果略好于 BPFA 算法，这说明增加了聚类特征的字典学习对字典生成的效果和稀疏表示的精度都有一定的提高。

实验还对其他标准灰度图进行降噪实验，比较结果见表 4.2。表中对应噪声标准差的三行分别是 K-SVD，BPFA 和 CLBP 三种算法的降噪结果，最好的降噪结果以粗体表示。从结果可以看出，噪音强度越高，本章的 CLBP 算法降噪效果与其他两种算法相比更好。算法对纹理信息比较多的图像处理效果更好，且能保留图像的细节信息，具有更高的峰值信噪比。

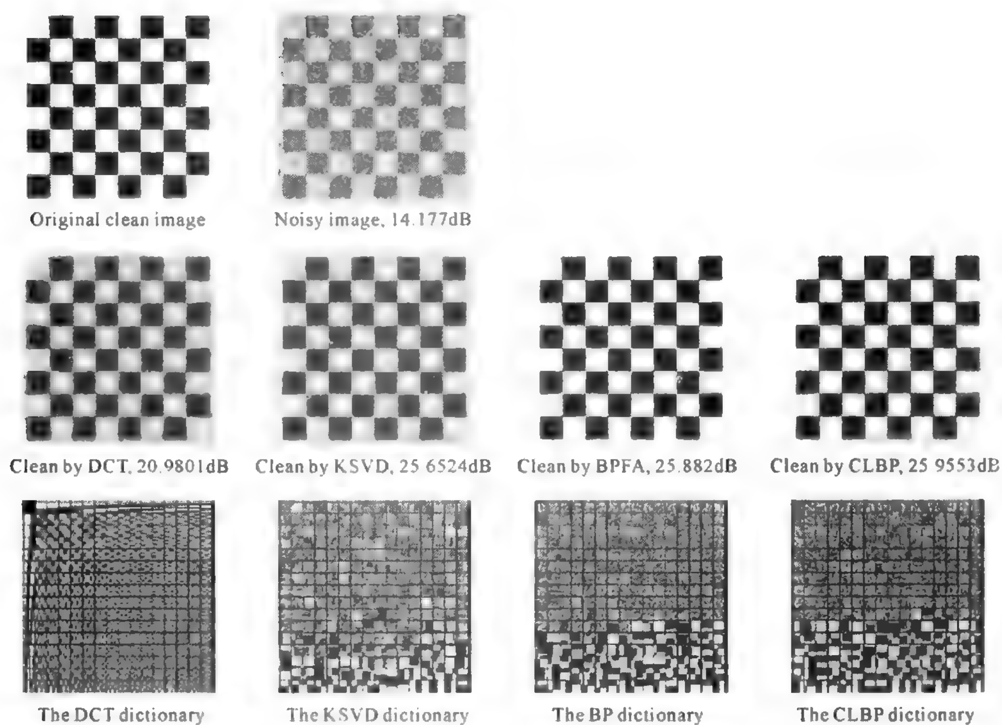


图 4.1　方形棋盘降噪效果比较，$\sigma = 50$

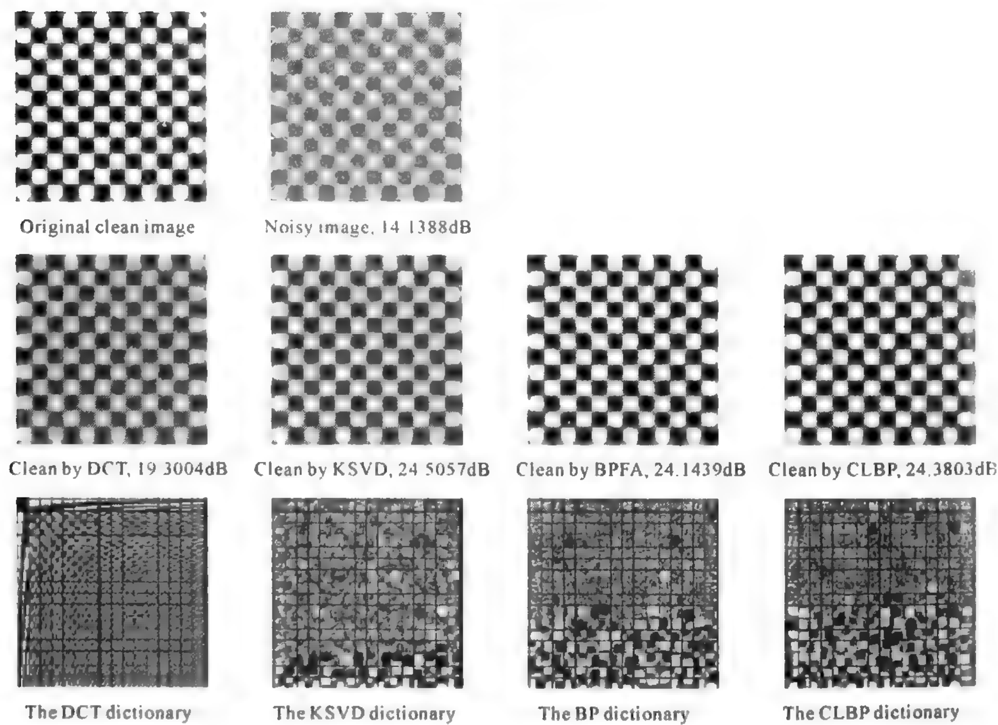


图 4.2　圆形棋盘降噪效果比较，$\sigma = 50$

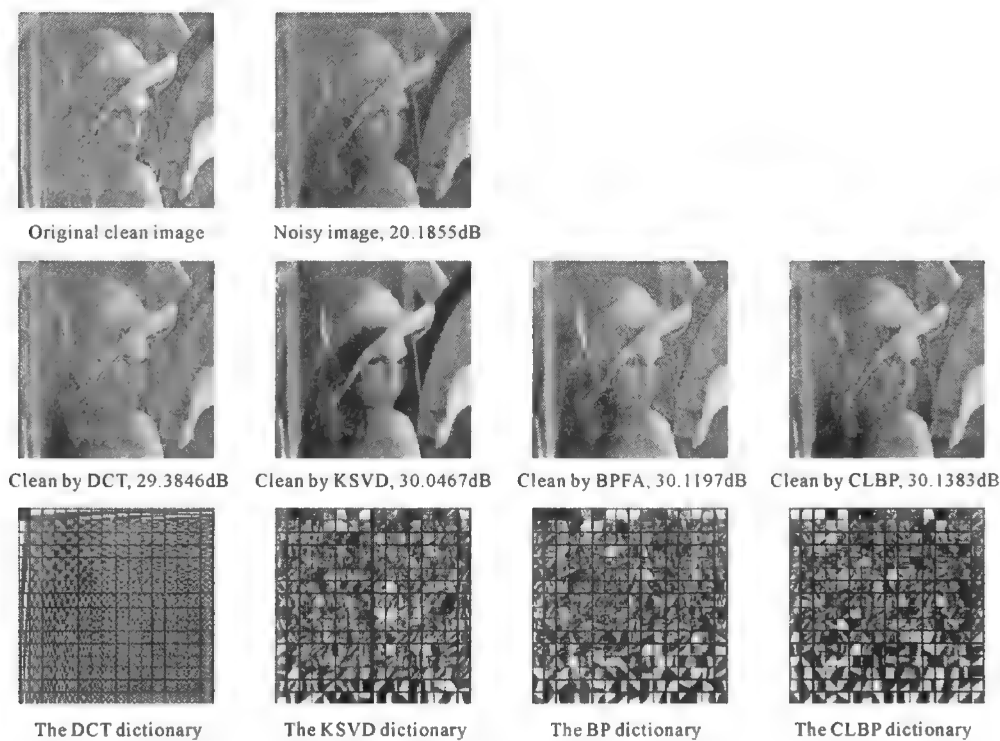


图 4.3　Lena 降噪效果比较，$\sigma = 25$

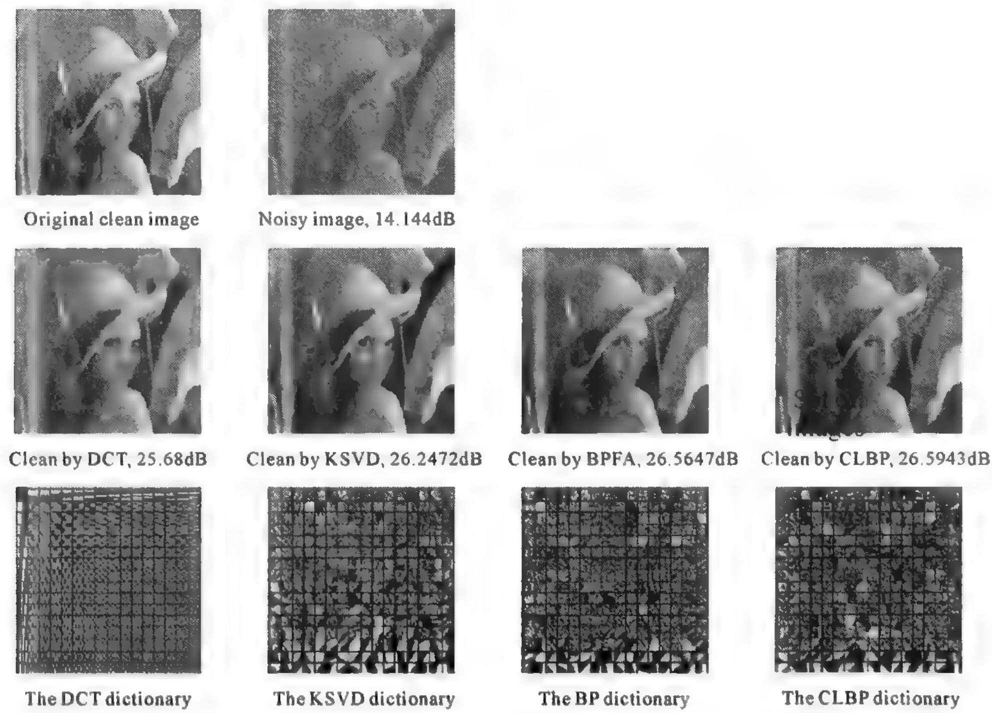


图 4.4　Lena 降噪效果比较，$\sigma = 50$

表 4.2　降噪效果比较

σ	Board	c. Board	House	Lena	Barbara	Peppers	Boat
10	**35.3288**	**33.5655**	35.7194	**34.9811**	**33.9648**	33.0361	**33.2668**
	33.3381	33.1077	**36.0069**	34.8058	33.5645	**33.1978**	33.1549
	32.9859	33.2175	35.8315	34.8241	33.4818	33.0838	33.1625
20	**31.5166**	**30.2185**	**32.4482**	31.1418	**30.2945**	29.0244	29.4800
	30.2249	29.4525	32.3550	31.2655	30.0986	29.3309	29.6189
	30.1296	29.4476	32.4017	**31.2802**	30.0553	**29.3580**	**29.6306**
30	**29.0106**	**28.3307**	29.709	29.0865	28.0772	26.8573	27.4420
	28.3914	26.9518	29.7898	29.1260	28.0947	27.0301	27.6281
	28.7617	27.1393	**29.8238**	**29.1423**	**28.1133**	**27.1599**	**27.6552**
40	26.8590	**26.2462**	27.6555	27.5303	26.5685	25.5140	26.1013
	27.2444	25.6265	28.0798	27.5876	26.8205	25.5732	26.2078
	27.3587	26.1133	**28.0854**	**27.6057**	**26.8336**	**25.6581**	**26.2222**
50	25.733	24.5511	26.3316	26.3341	25.3988	23.8883	25.0154
	26.0195	24.3371	26.6080	26.5857	25.7140	24.2172	25.1676
	26.0349	**24.5512**	**26.6689**	**26.5973**	**25.7428**	**24.2694**	**25.1815**
60	**24.5125**	**23.4861**	25.2107	25.1594	24.3645	22.8691	24.0950
	24.4457	22.8754	25.4840	25.5971	25.0072	23.1286	24.2708
	24.4837	23.0388	**25.5322**	**25.6105**	**25.0212**	**23.1294**	**24.3148**
70	23.286	**22.704**	24.0768	24.4672	23.2922	21.7251	23.2977
	23.5032	21.5209	24.3508	24.9796	24.1779	22.4899	23.6383
	23.5665	21.5942	**24.412**	**25.0071**	**24.2451**	**22.6081**	**23.6736**

4.5 小　结

稀疏降维在图像字典学习中有非常重要的应用，本章针对字典学习建模中存在的字典维度不能根据观测数据自适应调整的问题，以图像降噪为应用对象，给出了一种基于高斯过程聚类的贝叶斯非参数字典学习方法。基于高斯过程聚类的建模方法更适合图像数据的特点，能够使字典和稀疏表示模型在图像数据集上具有很好的推广性。在图像降噪过程中，由样本集自适应生成的列约束高斯字典作为对包含聚类特征的稀疏向量的稀疏表示测量矩阵，使字典学习的解具有一定的基于模型解释的可靠性，获得的字典和稀疏表示具有优化的解。实验结果表明该方法在模型精度、稀疏度和字典维度的自适应性上有优势。并且在图像降噪过程中，通过设定可调整参数可以控制降噪过程以获得更可靠的优化预测解。

第5章 基于狄利克雷过程的聚类方法

聚类分析是发现数据信息中存在的各种关系和规则，进行快速信息检索的有效途径，在模式识别、图像处理、计算机视觉、模糊控制等领域有广泛的应用。聚类将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程，它所生成的类的集合是一组数据对象的集合，同一个类中的对象彼此相似，与其他类中的对象却相异。一个好的聚类算法应能识别任意数据形态，对数据的输入顺序不敏感，随输入数据的大小线性扩展，当数据维数增加时也具有良好的可伸缩性。常用的分割聚类方法、层次聚类方法、基于密度的聚类方法、基于网格的聚类算法等在维数比较低的情况下能够生成质量较高的聚类结果，但不能适应高维数据特别是高维稀疏数据聚类。

传统的聚类算法可分以下五类：①划分方法，将数据集随机划分为 k 个子集，随后通过迭代重定位技术试图将数据对象从一个簇移到另一个簇来不断改进聚类的质量。②层次方法，对给定的数据对象集合进行层次的分解，根据层次的形成方法，又可以分为凝聚和分裂方法两大类。③基于密度的方法，根据领域对象的密度或者某种密度函数来生成聚类，使得每个类在给定范围的区域内必须至少包含一定数目的点。④基于网格的方法，将对象空间量化为有限数目的单元，形成一个网格结构，使所有聚类操作都在这个网格结构上进行，使聚类速度得到较大提高。⑤基于模型的方法，为每个类假定一个模型，寻找数据对给定模型的最佳拟合。

传统聚类方法是基于距离进行聚类的，数据间相似性的计算一般通过欧几里得距离、绝对值距离或明基考斯距离等方法进行。而这些聚类方法在高维数据集中进行聚类时，主要遇到两个问题：①高维数据集中存在大量无关的属性，使得在所有维中存在簇的可能性几乎为零。②高维空间中数据比低维空间中数据分布要稀疏，其中数据间距离几乎相等是普遍现象。对于高维稀疏数据，传统聚类方法很难反映数据之间的差异程度，从而难以得到正确的聚类结果。

本章首先对贝叶斯非参数方法实现聚类的过程进行分析，并对其进行应用验

证，给出适应背景剪除数据特点的狄利克雷聚类方法；再对高维稀疏聚类问题，以Pólya Tree为建模基础，给出一种高维稀疏数据聚类的方法，通过图像标注为应用和验证平台，证实算法的有效性。

5.1 贝叶斯非参数聚类

令$\boldsymbol{Y}\in\mathbf{R}^{N\times P}$是$P$个观测数据$\{\boldsymbol{y}_i\in\mathbf{R}^N\}_{i=1}^P$的集合，聚类的目标是寻找一个函数$\hat{f}:\mathbf{R}^N\to\{1,\cdots,K\}$，使得$\boldsymbol{Y}$中每个数据都映射到$K$簇中的一个。用概率的方法对聚类问题进行建模：

$$\hat{f}(\boldsymbol{y}_i)=\operatorname{argmax}\ \mathrm{P}(\boldsymbol{y}_i;\theta_k),\ k=1,\cdots,K \tag{5.1}$$

其中Φ_k是第k个簇的参数。例如，如果$\mathrm{P}(y_i;\theta_k)$是高斯函数，则θ_k是第k个高斯分布的均值和方差。如果采用贝叶斯方法，需要给θ_k赋予先验概率。

在有观测值$\boldsymbol{Y}$的条件下，根据贝叶斯公式，参数$\Theta=\{\theta_1,\cdots,\theta_K\}$的似然函数为：

$$\mathcal{L}(\Theta;\boldsymbol{y}_1,\cdots,\boldsymbol{y}_P)=\prod_{i=1}^{P}\sum_{k=1}^{K}P(C_i=k)P(\boldsymbol{y}_i\mid C_i=k;\Theta) \tag{5.2}$$

其中C_i表示数据$\boldsymbol{y}_i$所属的簇。直接通过最大似然计算每个参数的过程中，需要对所有的C_i求边缘分布，但这个边缘分布往往难于计算，所以需要寻找其他的计算方法，例如EM(Expectation Maximization algorithm)等。

簇的个数K对模型有很大的影响，如果K比实际的簇数小，则同一簇中数据的相似性降低，如果K过大，则形成孤立点的可能性增大。参数方法中事先假定簇的个数，降低了聚类的有效性和合理性。而非参数方法恰好可以避免这个问题，在贝叶斯非参数方法中，K的值由观测数据决定。观测数据的概率模型为：

$$p(\boldsymbol{y}_i)=\int_{\Theta}P(\boldsymbol{y}_i\mid\theta)G(\mathrm{d}\theta) \tag{5.3}$$

其中$\theta\in\Theta$，G是无限维函数空间中的一个未知的混合分布。把式(5.3)写为层次形式，得到：

$$\begin{aligned}\boldsymbol{y}_i\mid\theta_k&\sim p(\boldsymbol{y}_i\mid\theta_k)\\\theta_k\mid G&\sim G(\mathrm{d}\theta)\\G&\sim P(G)\end{aligned} \tag{5.4}$$

在贝叶斯非参数中，无限维空间中的G的先验$P(G)$通常赋以随机过程。随机过程在本书第2章中已经进行详细的描述，本处不作赘述。

5.1.1 基于狄利克雷过程的聚类

如果 G 由狄利克雷过程生成，记为 $G \sim \mathrm{DP}(\gamma, H)$，其中 γ 是收敛参数，H 是基础测度。假设聚类的空间以 Ω 表示，对于 Ω 的一个分割 $C_1, \cdots, C_k$ 有

$$(G(C_1), \cdots, G(C_k)) \sim \mathrm{Dir}(\gamma H(C_1), \cdots, \gamma H(C_k)) \tag{5.5}$$

其期望是 $\mathrm{E}[G(C_j)] = H(C_j)$，方差为 $\mathrm{Var}[G(C_j)] = \dfrac{H(C_j)(1 - H(C_j))}{\gamma + 1}$。对 G 进行边缘积分计算后，$\boldsymbol{y}_{1:P}$ 表现出聚类的效果。在已有 i 个观测数据 $\boldsymbol{y}_i$ 的已经聚类的条件下，对第 $i+1$ 个观测数据，它要么属于前 i 个数据构成的一个簇，要么构成一个新的簇，即

$$\boldsymbol{y}_{i+1} \mid \boldsymbol{y}_{1:i} \sim \frac{\gamma}{\gamma + i} H(\boldsymbol{y}_{i+1}) + \sum_{j=1}^{N_c} \frac{n_j}{\alpha + i} \delta_{\theta_j}(\boldsymbol{y}_{i+1}) \tag{5.6}$$

其中 N_C 是当前已有的簇的个数，n_j 是第 j 个簇中已有的观测数据的个数，θ_j 是第 j 个簇的参数。这个过程就是中国餐馆过程。关于中国餐馆过程的描述详见第2章。

图 5.1 显示了狄利克雷过程对一维数据的聚类结果，数据均依照高斯分布随机生成，均值分别选取 $-2, 2, 10$，方差为 0.5。三次实验的循环次数分别是 10，100，500，每行中左侧是先验分布，右侧是后验分布，灰色区域是根据先验或后验采样的范围，实线为估计的均值，虚线表示中间采样值。

图 5.2 显示狄利克雷过程对二维数据的聚类过程，分别是过程中循环 10，20，70，100 次的效果。

从上面两个图中可以看出，狄利克雷过程对数据的聚类是有效的，而且从对二维数据的聚类过程可以发现，聚类过程中 K 不是固定不变的，而是随着观测数据发生变化。

5.1.2 视频背景剪除中的狄利克雷过程聚类

背景剪除是视频处理中的一种常用方法，例如前景检测、目标跟踪等。背景剪除的过程一般是先建立背景模型，通过训练视频对模型参数进行学习；将当前的视频帧与背景模型进行比较，任何存在较大差异的区域都被认为是前景物体。

近年来，人们为解决背景剪除问题提出了许多方法，例如高斯模型、高斯混合模型、核密度估计等。高斯模型由 Wren 等人提出，模型对帧中每个像素分别建立相应的密度分布函数，尽管模型在室内场景中取得成功的应用，但对于复杂的室外场景，模型对于往复运动的背景物体识别率很低，例如，摇晃的树枝等。混合高斯模型用多个加权的高斯分布来描述每个像素，在处理新一帧图像时，如果某个像素

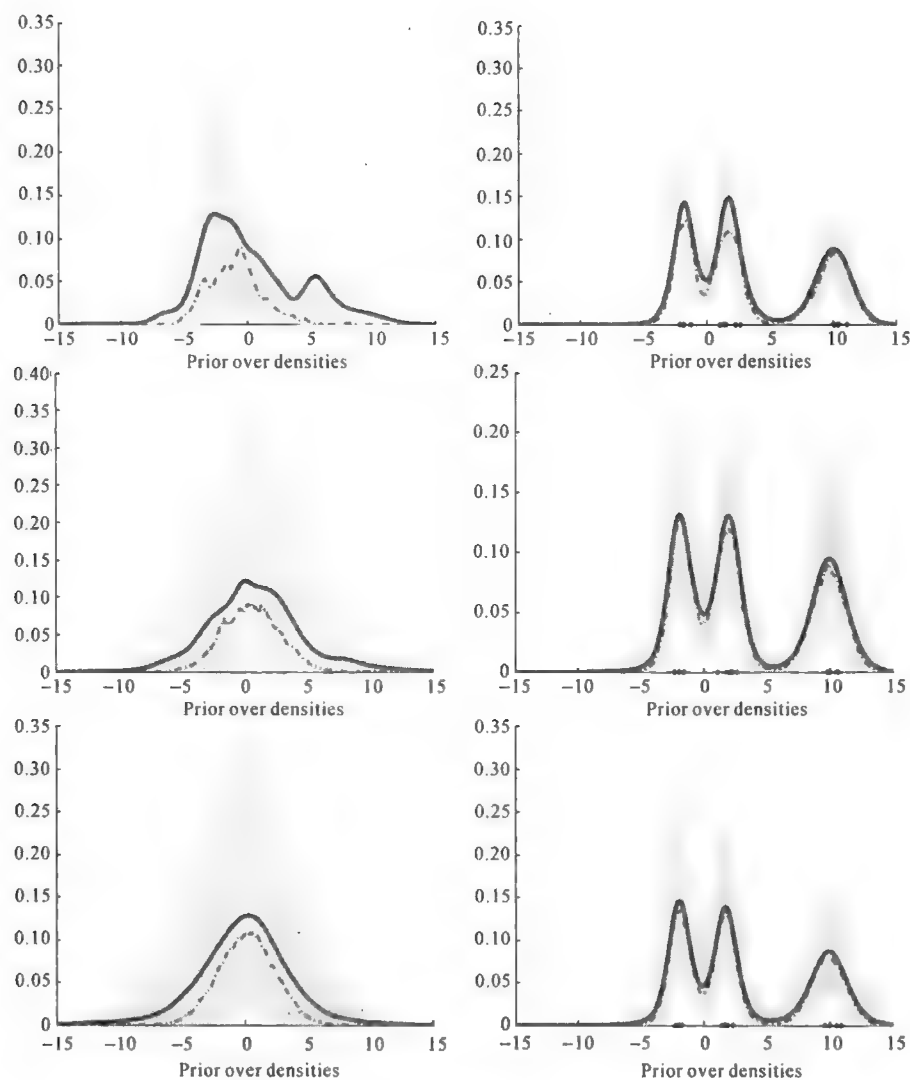


图 5.1　DP 对一维高斯数据的聚类

可以用混合高斯的背景模型描述，则认为此像素属于背景，否则将此像素分类为前景像素。Stauffer 等人给出了更新混合高斯模型参数的高效方法，但其高斯核的个数是固定的。Zoran 等人给出了高斯核自动选择的自适应 ADE 背景剪除算法。其他的背景剪除算法还包括 Graph-cut 算法、基于主成分分析的背景剪除算法等。我们基于贝叶斯非参数对聚类的自适应性，提出基于狄利克雷过程的背景剪除算法。

在背景剪除之前，首先将每一帧图像分为多个大小相同的块，将块中所有像素点的值映射，得到一个离散的集合：$\boldsymbol{d}(\vec{x}_i^{(t)}) = (d(\vec{x}_{i1}^{(t)}), d(\vec{x}_{i2}^{(t)}), \cdots)$。一种方法是利用直方图窗口进行分割，用大小为 W 的直方图窗口把视频中每一帧图像分为相同大小的块，以像素 i 为中心的窗口中所有像素的密度值得到 $h_i = (h_{i1}, \cdots, h_{iM})$。

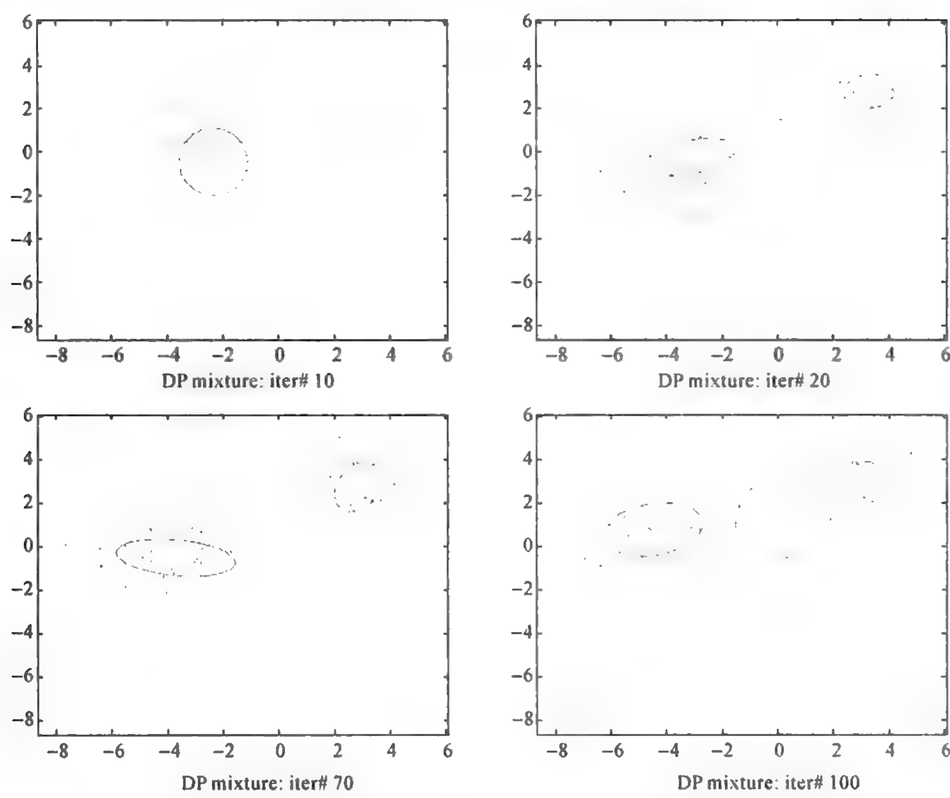


图 5.2 DP 对二维高斯数据的聚类

对于 $\boldsymbol{d}(\vec{x}_i^{(t)})$，令

$$\begin{aligned} F(\boldsymbol{d}(\vec{x}_i^{(t)}) \mid \theta_i) &= \frac{1}{Z(\boldsymbol{d}(\vec{x}_i^{(t)}))}\prod^{j}\theta_{ij}^{d(\vec{x}_{ij}^{(t)})} \\ &= \frac{1}{Z(\boldsymbol{d}(\vec{x}_i^{(t)}))}\exp(\sum_j d(\vec{x}_{ij}^{(t)})\log(\theta_{ij})) \end{aligned} \tag{5.7}$$

其中 $Z(\boldsymbol{d}(\vec{x}_i^{(t)}))$ 是标准化函数。令 Θ_i 服从狄利克雷分布

$$\begin{aligned} G_0(\theta_i \mid \beta\pi) &= \frac{1}{B(\beta\boldsymbol{\pi})}\prod_j \theta_{ij}^{\beta\pi_j-1} \\ &= \frac{1}{B(\beta\boldsymbol{\pi})}\exp(\sum_j(\beta\pi_j - 1)\log(\theta_{ij})) \end{aligned} \tag{5.8}$$

其中 $B(\beta\pi) = \dfrac{\prod_j \Gamma(\beta\pi_j)}{\Gamma(\beta)}$，$\beta$ 是正实数，$\boldsymbol{\pi}$ 是概率向量。

F,G_0 是共轭的，得到后验：

$$\begin{aligned} p(\theta_i \mid \boldsymbol{d}(\vec{x}_i^{(t)})) &\propto F(\boldsymbol{d}(\vec{x}_i^{(t)}) \mid \theta_i)G_0(\theta_i) \\ &\propto \exp(\sum_j(d(\vec{x}_{ij}^{(t)}) + \beta\pi_j - 1)\log(\theta_{ij})) \\ &= G_0(\theta_i \mid \boldsymbol{d}(\vec{x}_i^{(t)}) + \beta\boldsymbol{\pi}) \end{aligned} \tag{5.9}$$

则生成新簇的概率为：

$$q_{i0} \propto \alpha \int_{\theta_i} F(\boldsymbol{d}(\vec{x}_i^{(t)}) \mid \theta_i) G_0(\theta_i) \mathrm{d}\theta_i = \frac{\alpha B(\boldsymbol{d}(\vec{x}_i^{(t)}) + \beta \boldsymbol{\pi})}{Z(\boldsymbol{d}(\vec{x}_i^{(t)})) B(\beta \boldsymbol{\pi})} \tag{5.10}$$

属于已有的第 k 个簇的概率为：

$$q_{ik} \propto n_k^{-i} F(\boldsymbol{d}(\vec{x}_i^{(t)}) \mid \theta_k^*) = \frac{n_k^{-i}}{Z(\boldsymbol{d}(\vec{x}_i^{(t)}))} \exp(\sum_j d(\vec{x}_{ij}^{(t)}) \log(\theta_{ij})) \tag{5.11}$$

$k = 1, \cdots, N_c$。更新簇参数 θ_k^* ：

$$\begin{aligned} \theta_k^* &\sim G_0(\theta_k^*) \prod_{i|s_i=k} F(\boldsymbol{d}(\vec{x}_i^{(t)}) \mid \theta_k^*) \\ &\propto \exp(\sum_j (\beta\pi_j + \sum_{i|s_i=k} d(\vec{x}_{ij}^{(t)}) - 1) \log(\theta_k^*)) \\ &\propto G_0(\theta_k^* \mid \beta\pi + \sum_{i|s_i=k} \boldsymbol{d}(\vec{x}_i^{(t)})) \end{aligned} \tag{5.12}$$

具体算法如表 5.1 所示。

表 5.1　BSMDP 算法

算法 3：BSMDP 算法
输入：β、π 和 $\{\vec{x}_1^{(t)}, \ldots, \vec{x}_N^{(t)}\}$，其中 $t = 0, \ldots, T-1$ 输出：S (1) $d(\vec{x}_i^{(t)}) = f_{map}(\vec{x}_i^{(t)})$ ； (2)当 $t = 0$，$\theta_{i0}^* \sim G_0(\theta_{i0}^*) \prod F(\mathrm{d}(\vec{x}_i^{(0)}) \mid \theta_{i0}^*)$； (3)当 $t < T$，对每个 i，计算： $q_{i0} \propto \frac{\alpha B(\boldsymbol{d}(\vec{x}_i^{(t)}) + \beta\pi)}{Z(\boldsymbol{d}(\vec{x}_i^{(t)})) B(\beta\pi)}$ ； $q_{ik} \propto \frac{n_k^{-i}}{Z(\boldsymbol{d}(\vec{x}_i^{(t)}))} \exp(\sum_j d(\vec{x}_{ij}^{(t)}) \log(\theta_{ij}))$ ； (4)根据 $q_{i0}, \cdots, q_{iN_c}$ 得到 k ； (5)如果 $k \in \{1, \ldots, N_c\}$，则 $S_i \leftarrow k$，否则 $N_c \leftarrow N_c + 1$，$S_i \leftarrow N_c$ ； (6)更新 θ：$\theta_{ik}^* \sim G_0(\theta_{ik}^*) \prod_{i\|S_1=k} F(\mathrm{d}(\vec{x}_i^{(0)}) \mid \theta_{ik}^*)$； (7) 迭代(3)～(6)直至收敛。

5.1.3 实验结果与分析

为了验证本算法的有效性，我们以算法对 Campus 视频进行背景剪除处理。在实验中，对于块尺寸 M，需要根据视频的复杂度进行设定。M 值越大，计算速度越快，但同时会丢失更多的细节信息。图 5.3 显示了算法的背景剪除效果。第一行是视频中的原始图像，分别选取视频中第 900 帧、1200 帧和 2500 帧。第二行是通过算法背景剪除后的图像，第三行是从视频中学习的背景图像，最后一行是当前视频帧与学习背景的差异图。从图中可以看出，算法能够学习出视频背景，但效果不够理想。

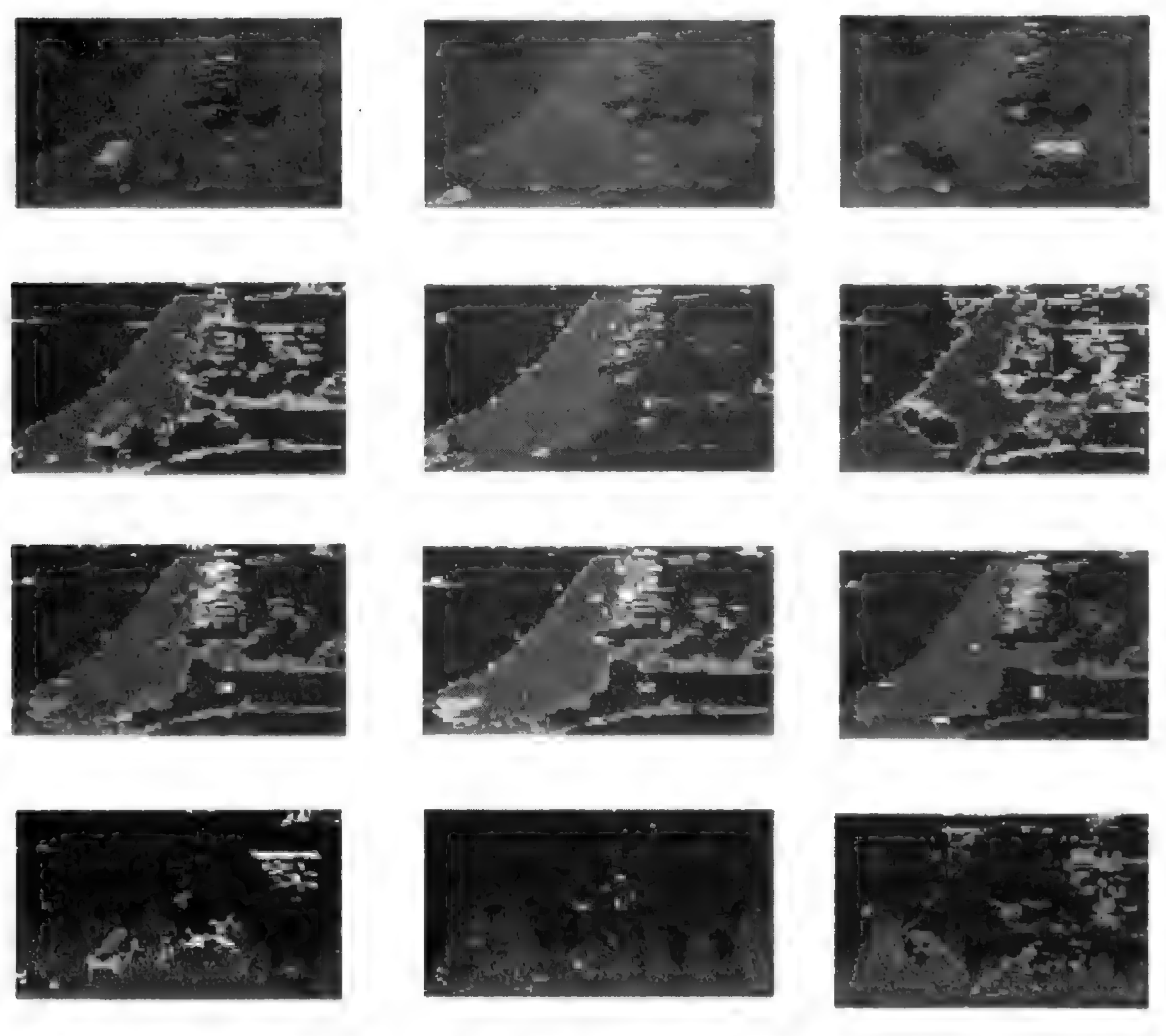


图 5.3　背景剪除实验结果

5.2 基于 Pólya Tree 的高维稀疏聚类

上一节中讨论了贝叶斯非参数狄利克雷过程对聚类的实现，并在此基础上给出适应背景剪除数据特点的狄利克雷过程聚类方法。本节对高维稀疏聚类的贝叶斯非参数方法进行研究，提出基于 Pólya Tree 的高维稀疏聚类方法，并利用图像标注实验对其进行验证。

5.2.1 高维稀疏聚类问题和现有方法

目前，高维稀疏数据多指高属性维稀疏数据，即假设有 N 个数据，每个数据有 M 个属性，M 的值较大，且每个数据的大部分属性值为零。高维稀疏数据的产生与生物信息学的发展和电子信息化的加深密不可分，在实际的高维数据应用中，往往需要对某类具有上百个属性的对象进行聚类，从而很难得到理想的聚类结果。至今，有很多文献对如何进行高维对象之间的聚类进行了研究，提出的方法主要包括频繁模式挖掘、特征转换法、特征选择/子空间聚类等。

1)频繁模式挖掘方法

频繁模式挖掘源自关联分析，确定关联规则中的频繁项集和它们的支持度问题，被称之为频繁模式的挖掘。对于高维稀疏数据集，频繁模式挖掘算法可用来发现有共同调控关系的属性或属性组，基于频繁模式的关联规则可以用来构建属性网络。频繁模式的挖掘算法可以划分为三类：基于特征计数的算法、基于行计数的算法和混合计数算法。基于特征计数的算法有 A-close，CLOSET，MAFIA，CHARM 和 CLOSET+ 等。这些算法分别采用宽度优先搜索(BFS)和深度优先搜索(DFS)算法对搜索特征计数树从根部进行搜索，保证所有的特征组合都访问到。基于行计数的算法有 Carpenter，FARMER，TOPKERS，TD-Close 等。混合计数算法有 COBBLER 等。

2)特征转换法

特征转换法是高维聚类常用的一种方法，这种方法先将原数据进行降维处理，然后在降维的空间中进行聚类。最普遍的是通过主成分分析(PCA)把数据映射到一个低维的子空间中，这个空间可以保留数据的多数变量，然后在这个低维的子空间中用欧几里得距离构建相似矩阵，从而得到更精确的聚类。但是，如果数据在这个子空间中的投影是非线性的，这种算法的效果很差，例如，图 5.4 显示的降维投影，数据(用“*”表示)经过 PCA 降维后的投影用虚线表示，合理的降维结果用实线表示。左图的合理投影是非线性的。右图是线性的。Yeung 等证实了 PCA 作

为降维的方法对基因表达数据的聚类是不合适的。

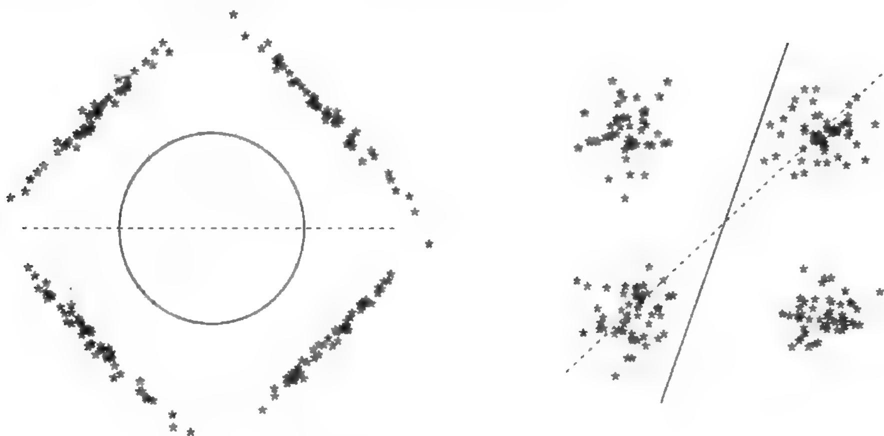


图 5.4　主成分分析降维

谱聚类(Spectral clustering)是另一个常用的技术,它是一个通过对相似矩阵以 Laplacian Eigenmap 的降维方式降维之后,再执行 K-means 的一个过程。在这个算法中,相似矩阵通过核函数计算。常用的核函数是高斯核,通过对相似矩阵的 Laplacian 矩阵求解特征值和特征向量,将特征值从大到小进行排列,取最大的 k 个特征值对应的特征向量组成矩阵,再对这个特征向量矩阵用 K-means 算法进行聚类,结果中每一行所属的类别就是数据点分别所属的类别。尽管特征转换法的多种算法针对不同的数据集展现了良好的聚类效果,但该方法在执行中,一方面难于确定合适的 k 值,另一方面,高维空间中存在大量无关维而掩盖了簇,给聚类造成困难,同时,在聚类过程中容易产生无意义的簇。因此,该方法适合事先已知大多数维都相关的高维数据集进行聚类。

3)特征选择/子空间法

特征选择只在那些相关的子空间上执行挖掘任务,因此它比特征转换能更有效地减少维。特征选择一般使用贪心策略等搜索方法搜索不同的特征子空间,然后使用一些标准来评价这些子空间,从而找到所需的簇。

子空间聚类方法在相同数据集的不同子空间上发现聚类,这些子空间通常要尽可能多地包含原始数据的特征。由于基于不同的子空间进行聚类,该方法需要使用一定的测评标准来筛选出需要聚类的簇,从而在多个子空间聚类的结果中选择能够使测评标准最大(或最小)的那个子空间作为算法的聚类结果。另外,和特征选择一样,子空间聚类需要使用一种搜索策略,选择的搜索策略对聚类结果有很大的影响。根据搜索方向的不同,可以将子空间聚类方法分成两大类:自顶向下的搜索策略和自底向上的搜索策略。CLIQUE,ENCLUS 算法采用了自底向上的搜

索策略，它们的改进算法 MNFIA，LTREE，CBF，DOC 都采用某种策略动态查找最佳分割点，以获得比较稳定的结果。然而，自底向上的策略很容易导致有重叠的簇产生，即某些点不属于一个簇或属于多个簇。该类方法一般都需要两个参数：网格的大小和密度阈值。两个参数的值对最后形成簇的质量有很大影响，但是要确定它们非常困难。自顶向下算法为数据的每个部分都建立簇，这意味着不会有重复的簇产生，一个点只能赋予一个簇。许多算法也产生一个集合来分析孤立点。PROCLUS 是最典型的自顶向下算法。但这种方法中的参数，例如，簇的数量、相同或相近的簇的大小很难确定。另外，子空间聚类的最大挑战在于如何找到最恰当的子空间，这也阻止了该方法的更多应用和发展。

另一种直观的聚类方法是基于高斯混合模型的聚类，但对于高维稀疏数据，协方差矩阵通常是非奇异的。为了使高斯混合模型适应与高维稀疏数据的聚类，需要对协方差矩阵进行正则化处理。

2009 年以来，随着贝叶斯非参数方法的快速发展，尤其在 2011 年，贝叶斯非参数方法的研究者们提出多种基于贝叶斯非参数方法的高维稀疏聚类，例如，Yau 等提出的以高斯过程为基础的“K-成分层次贝叶斯高斯混合”模型，Socher 等人提出的“谱中国餐馆过程”，Nia 提出的“高斯和非对称 Laplace 聚类模型”，Adams 等人在 2010 年 NIPS 会议上发表的以树结构 Stick-breaking 过程对层次数据进行建模的方法，等等。

在分析了现有方法的基础上，本章给出一种新的对高维稀疏数据聚类的方法，与树结构的 Stick-breaking 过程类似，我们也采用树的结构对高维稀疏数据赋予先验，但与 Stick-breaking 树是多叉树不同，新方法构造二叉树，利用二叉树存储和遍历的优点，实现数据的快速聚类。

5.2.2 Pólya Tree

Pólya Tree 由 Pólya Urn 机制发展而来。正如第 2 章中对 Pólya Urn 机制的分析，它是最简单、最实用的生成可交换随机变量序列 $Y_1, Y_2, \cdots$ 的方法，且这些变量的值包含在有限集合 $E = \{0, \cdots, k\}$ 中。Pólya Tree 定义在 $E^* = \bigcup E$ 之上，从不同的罐中取球从而生成随机变量序列，Mauldin 等人证明了这些随机变量序列也是可交换的。Pólya Tree 分布为随机概率测度的分割定义了有限维度的分布，从而扩展了狄利克雷过程的思想。Pólya Tree 具体定义如下：

定义 5.1：(Pólya Tree)对于 $\mathbf{R}$ 上的随机概率测度 P，如果存在随机变量 $\boldsymbol{Y} = \{Y_0, Y_{00}, Y_{10}, \cdots\}$ 满足下列三个条件，则称 P 有 Pólya Tree 分布，分布的参数为 (Π, A)，记为 $P \sim \mathrm{PT}(\Pi, A)$。

(1) $\boldsymbol{Y}$ 中的随机变量是相互独立的；

(2) 对每个 $\varepsilon \in E^*$，都有 $Y_{\varepsilon 0} \sim \mathrm{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 0})$；

(3) 对每个 $m = 1,2,\cdots$ 和每个 $\varepsilon \in E^m$，

$$P(B_\varepsilon) = \left[\prod_{\substack{j=1 \\ \{\varepsilon_j = 0\}}}^{m} Y_{\varepsilon 1 \cdots \varepsilon_{j-1} 0}\right]\left[\prod_{\substack{j=1 \\ \{\varepsilon_j = 1\}}}^{m} 1 - Y_{\varepsilon 1 \cdots \varepsilon_{j-1} 0}\right]$$

其中，$E^m = \{\varepsilon = \varepsilon_1, \cdots, \varepsilon_m, \varepsilon_k \in \{0,1\}\}$，$E^0 = \{\phi\}$，$E^* = \bigcup_{m=1}^{\infty} E^m$。

对于 Pólya Tree，可以通过如下过程描述：

(1) 令 $\{B_0, B_1\}$ 是对 $\mathbf{R}$ 的一个测度分割，$\{B_{00}, B_{01}\}$ 是对 B_0 的分割，$\{B_{10}, B_{11}\}$ 是 B_1 的分割，重复这个过程。在第 m 阶分割，$B_\varepsilon, \varepsilon \in E^m$ 被分割为 $\{B_{\varepsilon 0}, B_{\varepsilon 1}\}$。在分割中，允许 $B_{\varepsilon 0} = \phi, B_{\varepsilon 1} = B_\varepsilon$ 情况的存在。图 5.5 描述了这个分割过程。

<table>
<tr><td>Stage0</td><td colspan="12">$\mathbb{R}$</td></tr>
<tr><td>Stage1</td><td colspan="5">B_0</td><td colspan="7">B_1</td></tr>
<tr><td>Stage2</td><td colspan="2">B_{00}</td><td colspan="3">B_{01}</td><td colspan="4">B_{10}</td><td colspan="3">B_{11}</td></tr>
<tr><td>Stage3</td><td colspan="2">B_{000}</td><td>B_{010}</td><td colspan="2">B_{011}</td><td colspan="2">B_{100}</td><td colspan="2">B_{101}</td><td colspan="2">B_{110}</td><td>B_{111}</td></tr>
<tr><td>Stage4</td><td>B_{0000}</td><td>B_{0001}</td><td>B_{0101}</td><td>B_{0110}</td><td>B_{0111}</td><td>B_{1000}</td><td>B_{1001}</td><td>B_{1010}</td><td>B_{1011}</td><td>B_{1100}</td><td>B_{1111}</td><td>B_{1110}</td></tr>
</table>

…… …… …… …… ……

图 5.5 Pólya Tree 对测度空间的分割过程

(2) 定义随机变量序列 $\mathbf{Y} = \{Y_0, Y_{00}, Y_{10}, \cdots\}$ 和一个非负的实数参数序列 $A = \{\alpha_0, \alpha_1, \alpha_{00}, \alpha_{10}, \cdots\}$，对每个 $\varepsilon \in E^*$，有 $Y_{\varepsilon 0} \sim \mathrm{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1})$。$\phi 0$ 即为 0，$\phi 1$ 即为 1。

(3) $\mathbf{Y}$ 中的随机变量决定 P 中的条件分布 $Y_{\varepsilon 0} = P(B_{\varepsilon 0} \mid B_\varepsilon)$，换个角度讲，$P(B_{\varepsilon 0} \mid B_\varepsilon)$ 上信息的权重由参数 $\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}$ 决定。

如果对于任意 $\varepsilon \in E^*$，都有 $\alpha_\varepsilon = \alpha_{\varepsilon 0} + \alpha_{\varepsilon 1}$，则 Pólya Tree 是狄利克雷过程。

5.2.3 基于 Pólya Tree 的高维稀疏聚类

在分析了高维稀疏数据特征和 Pólya Tree 特点的基础上，方法采取在线(Online)的方式对观测数据进行聚类。这种选择一方面基于在线算法的计算量小，只根据已有观测值获得的结果和当前观测值进行计算，另一方面考虑到处理的数据维度高，批处理算法中计算开销大。图 5.6 给出了方法的整体过程，方法的输入可以是任何形态的数据集合，例如，文本文档、手写数字、图像等等。对于这些数据，首先计算其特征矩阵。例如，对于图像，通过第 4 章的字典学习方法，获得同类图像的字典，并将待聚类图像通过字典映射为特征矩阵。再对特征矩阵根据 Pólya Tree 聚类算法得到数据的聚类结果。

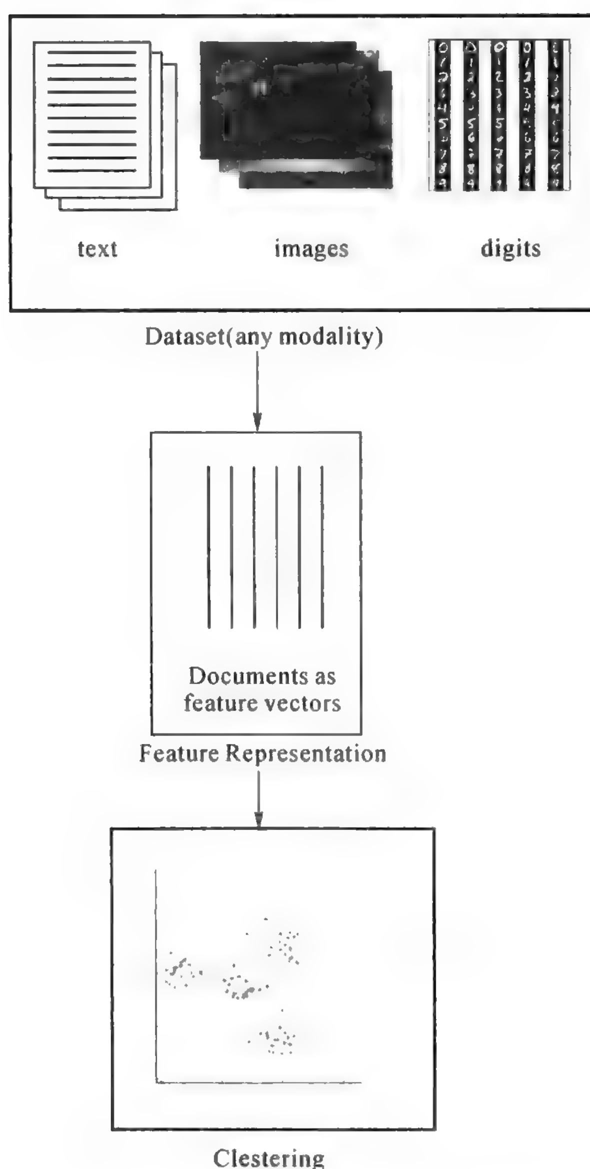


图 5.6　基于 Pólya Tree 聚类方法的框架

把 Pólya Tree 生成狄利克雷过程的构建过程看作是生成无穷分割，对于高维稀疏数据 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots\boldsymbol{x}_N\}$ 的聚类过程，有随机测度 $P\sim \mathrm{PT}(\Pi,A)$，存在随机变量 $\boldsymbol{\theta}=\{\theta_1,\theta_2,\theta_3,\cdots\}$ 满足 Pólya Tree 的条件，聚类过程为：

$$
\begin{aligned}
c_k &\sim \mathrm{Multinomial}(\pi_k) \\
\pi_k &\sim \prod_{d=1}^{D}(\theta_{\varepsilon_d})^{x_{nd}}(1-\theta_{\varepsilon_d})^{1-x_{nd}} \\
\theta_{\varepsilon_d} &\sim \mathrm{Beta}(\alpha_{\varepsilon_{d0}},\alpha_{\varepsilon_{d1}})
\end{aligned}
\tag{5.13}
$$

其中 D 是 Pólya Tree 的当前深度，ε_d 是聚类 Pólya Tree 中深度为 d 的一个节点，其左右两个子节点分别是 ε_{d0} 和 ε_{d1}，相应的参数为 $\theta_{\varepsilon_{d0}}$ 和 $\theta_{\varepsilon_{d1}}$，$\alpha_{\varepsilon_d}=\alpha_{\varepsilon_{d0}}+\alpha_{\varepsilon_{d1}}$。

树的深度 D 随着观测数据的增加发生变化，这也是用 Pólya Tree 对聚类建模的优点。聚类的簇的范围受 λ 参数控制，当节点 ε_d 所属簇的原子不断增加，当原子的方差超过 λ，则将该簇分割为两个子簇，即对节点 ε_d 分裂为两个子节点，$D=D+1$。对该簇中的原子，根据 x_{d+1} 确定其归属于左子簇还是右子簇。

在此过程中，对高维稀疏数据 $\boldsymbol{X}$ 的聚类过程也是 Pólya Tree 的生成过程。由于 Pólya Tree，其后验也是 Pólya Tree，由观测数据 $\boldsymbol{x}_i$ 更新的 Pólya Tree。对于观测稀疏向量 $\boldsymbol{x}_i$，如果它落入 Pólya Tree 中节点 ε_d 对应的簇，则 Pólya Tree 更新的参数 $\alpha_{Path_{\varepsilon_d}}(\boldsymbol{x}_i)=\alpha_{Path_{\varepsilon_d}}(\boldsymbol{x}_i)+1$，其中 $Path_{\varepsilon_d}$ 是节点 ε_d 及其所有祖先节点。

如果 $P\sim \mathrm{PT}(\Pi,A)$，P 的后验也是 Pólya Tree。在有观测数据 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N$ 条件下，

$$P\mid\{\boldsymbol{x}_i\}\sim \mathrm{PT}(\Pi,A(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)) \tag{5.14}$$

其中 $\alpha_\varepsilon(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)=\alpha_\varepsilon+n_\varepsilon$，$n_\varepsilon$ 是 $\{\boldsymbol{x}_i\}$ 落入 ε 节点对应簇的数据个数，则 θ_{ε_d} 的更新为：

$$\theta_{\varepsilon_d}\sim \mathrm{Beta}(\alpha_{\varepsilon_{d0}}+n_{\varepsilon_{d0}},\alpha_{\varepsilon_{d1}}+n_{\varepsilon_{d1}}) \tag{5.15}$$

具体算法 PT-HDSC 如表 5.2 所示。

表 5.2 PT-HDSC 算法

算法 4：PT-HDSC 算法
输入：$\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$ 输出：C (1)根据父节点 ε_0 和参数 θ_{ε_0} 初始化 Pólya Tree； (2)第 i 次迭代，第 1 步：得到 x_i 在 Pólya Tree 中的路径； (3)第 2 步：根据式(5.15)更新路径中的 θ； (4)第 3 步：得到路径中最深层节点 ε_d，从而计算 c_{ε_d}； (5)如果 $c_{\varepsilon_d}>\lambda$，则初始化参数 ε_{d0} 和 ε_{d1}，并将 c_{ε_d} 的节点分类到 $c_{\varepsilon_{d0}}$ 和 $c_{\varepsilon_{d1}}$，更新参数 ε_{d0} 和 ε_{d1}； (6) 迭代(2)～(5)直至收敛。

5.2.4 实验结果与分析

为了验证基于 Pólya Tree 聚类方法的有效性，我们在 CIFAR 图像数据集上进行实验，该数据集包含 50 000幅训练图像和 10 000幅测试图像，图像包含 100 类物体，每幅图像均为 32×32×3 彩色图像。数据集中包含的图像的拍摄尺度、视

角、光线和背景都各不相同，这加大了图像识别和聚类的难度。

实验运行环境为四核 i5 280GHz 处理器，8GB 内存，matlab 版本为 R2010b。受实验条件限制，实验选取部分训练图像和测试图像进行实验。实验首先采用第 4 章字典学习的方法提取图像特征，并对字典的维度设置上限为 256，对于那些有效维度小于上限的字典，以 0 对其补全，从而得到维度一致的特征表示 $\boldsymbol{x}_i \in \mathbf{R}^{256}$。然后再通过 PT-HDSC 算法对其进行聚类，部分聚类结果如图 5.7 所示。

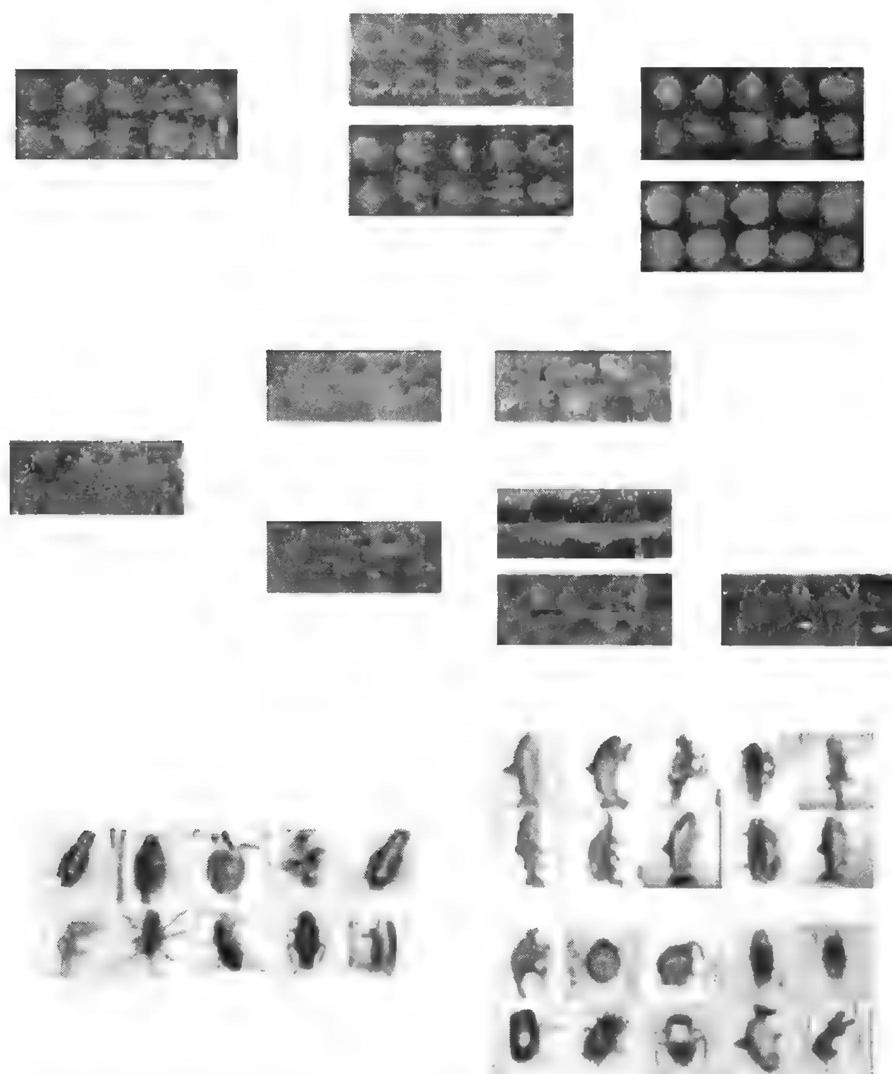


图 5.7　部分聚类结果

5.3 小　结

聚类分析是发现数据信息中存在的各种关系和规则，进行快速信息检索的有效途径。本章在分析贝叶斯非参数方法实现聚类的基础上，针对视频图像数据和高维稀疏数据的聚类问题进行研究。为了适应视频图像数据的聚类，提出基于混合狄利克雷过程的背景剪除方法，并通过实验证实了方法的可行性和有效性。继而，针对高维稀疏数据的特征，给出基于 Pólya Tree 的聚类方法。该方法对于待聚类的数据先通过特征提取，得到数据在高维属性下的稀疏表示，再对这些高维稀疏数据以 Pólya Tree 建模，得到能够适应数据增加树的深度、改变聚类数的一种聚类方法，并通过对 CIFAR 图像数据集的聚类实验证实了方法的有效性。

第6章
结束语

统计稀疏学习是计算机科学、统计学和认知科学的交叉领域，是一个新兴的统计学习研究方向，而贝叶斯非参数方法对统计稀疏学习中问题的研究有重要的作用。本书对贝叶斯非参数方法的构造方法、表达能力和推理机制进行了研究和讨论。在此基础上，研究了贝叶斯非参数方法对统计稀疏学习中稀疏表示、稀疏建模和稀疏降维问题的建模方法和推理过程，并将其应用于具体视觉任务，例如手写数字识别、图像降噪、视频背景剪除等，同时也利用这些视觉任务验证了方法的可行性和有效性。

贝叶斯非参数方法是表示和分析不确定性知识的有效工具，然而贝叶斯非参数方法的应用研究在国内尚处于起步阶段。本书基于贝叶斯非参数的统计稀疏表示、学习和推理，针对统计稀疏学习理论的主要问题，在分析贝叶斯非参数方法中典型的模型构建、学习方法和推理机制的基础上，对贝叶斯非参数方法对稀疏的表达能力、构建方法和推理机制进行了探索性研究。

针对统计稀疏学习中稀疏表达的建模问题，通过扩展稀疏向量的函数形式，提出自适应稀疏向量线性表达的贝叶斯非参数方法以获得更稀疏的稀疏表示模型。利用混合贝努利-贝塔过程，自动根据观测数据在已知测量矩阵上的稀疏投影频率调整稀疏向量的稀疏度。同时以高斯分布近似的拉普拉斯先验对 l_0 范数进行逼近，降低推理的复杂度和提高计算速度，并分别在人工单位脉冲数据集和手写数字识别数据集上证明了方法能够降低误差，提高识别率。

稀疏降维在图像字典学习中有非常重要的应用，本书针对字典学习建模中存在的字典维度不能根据观测数据自适应调整的问题，以图像降噪为应用对象，提出了一种基于高斯过程聚类的贝叶斯非参数字典学习方法。基于高斯过程聚类的建模方法更适合图像数据的特点，能够使字典和稀疏表示模型在图像数据集上具有很好的推广性。在图像降噪过程中，由样本集自适应生成的列约束高斯字典作为对包含聚类特征的稀疏向量的稀疏表示测量矩阵，使字典学习的解具有一定的基

于模型解释的可靠性，获得的字典和稀疏表示具有优化的解。实验结果表明该方法在模型精度、稀疏度和字典维度的自适应性上有优势。并且在图像降噪过程中，通过设定可调整参数可以控制降噪过程，以获得更可靠的优化预测解。

聚类分析是发现数据信息中存在的各种关系和规则，进行快速信息检索的有效途径。本书在分析贝叶斯非参数方法实现聚类的基础上，针对视频图像数据和高维稀疏数据的聚类问题进行研究。为了适应视频图像数据的聚类，提出基于混合狄利克雷过程的背景剪除方法，并通过实验证实了方法的可行性和有效性。继而，针对高维稀疏数据的特征，提出基于 Pólya Tree 的聚类方法。该方法对于待聚类的数据先通过特征提取，得到数据在高维属性下的稀疏表示，再对这些高维稀疏数据以 Pólya Tree 建模，得到能够适应数据增加树的深度、改变聚类数的一种聚类方法，并通过对 CIFAR 图像数据集的聚类实验证实了方法的有效性。

贝叶斯非参数方法是机器学习领域研究的主流和热点，它以坚实的统计与概率科学为基础，激发了机器学习领域新的研究主题。贝叶斯非参数方法在统计稀疏学习中将有更加深远的影响和广泛的作为。随着统计稀疏学习理论的发展，需要进一步研究与之相适应的贝叶斯非参数方法，以期从贝叶斯数据分析角度，利用非参数方法的不确定性表达能力，给出新的模型和方法。由于稀疏表达在视觉认知上的理论基础，采用视觉应用来验证提出的统计稀疏学习方法具有自然的动机并容易得到直观的结果。目前典型的视觉应用包括基于稀疏字典学习的图像降噪、图像补充、人脸识别、动作分割等应用。

参考文献

[1] 朱淼良.计算机视觉[M].杭州:浙江大学出版社,1997.

[2] 石光明,刘丹华.压缩感知理论及其研究进展[J].电子学报,2009, 37(5):1070—1081.

[3] 郭志波,杨静宇,刘华军,等.基于矩阵完备投影的快速主分量分析算法[J].中国图象图形学报, 2007, 12(4):628—632.

[4] 蔡泽民,赖剑煌.一种基于超完备字典学习的图像去噪方法[J].电子学报,2009, 37(2):347—350.

[5] 杨谦,齐翔林,汪云九.视皮层 V1 区简单细胞的稀疏编码策略[J].计算物理,2001, 18(2):136—143.

[6] 李清勇,胡宏,施智平,等.基于纹理语义特征的图像检索研究[J].计算机学报, 2006, 29(1):116—123.

[7] 侯彪,刘芳.基于脊波变换的直线特征检测[J].中国科学:E辑,2003, 33(1):65—73.

[8] H. Trevor, Tibshirani R, Jerome F. The Elements of Statistical Learning: Data Mining, Inference, and Prediction[M]. Berlin: Springer, 2009.

[9] Tibshirani R. Regression shrinkage and selection via the Lasso[J]. Journal of Royal Statist, 1996, 58(1):267—288.

[10] Donoho D L. Compressed sensing[J]. IEEE Transactions on Information Theory, 2006, 52(4):1289—1306.

[11] Attneave F. Some informational aspects of visual perception [J]. Psychological Review, 1954, 61 (3):183—193.

[12] Olshausen B A, Field D J. Sparse coding with an overcomplete basis set: a strategy employed by V1[J]. Vision Research, 1997, 37(23):3311—3325.

[13] Kay K N, Naselaris T, Prenger R J, et al. Identifying natural images from

human brain activity[J]. Nature, 2008(452):352—355.

[14] Baraniuk R G. Compressive sensing[J]. Lecture Notes in IEEE Signal Processing Magazine, 2007, 24(4):118—120.

[15] Mallat S, Zhang Z. Matching pursuits with time—frequency dictionaries[J]. IEEE Trans. Signal Processing, 1993, 41(12):3397—3415.

[16] Chen S, Donoho D, Saunders M. Atomic decomposition by basis pursuit[J]. SIAM Journal on Scientific Computing, 1999(20):33—61.

[17] Aharon M, Elad E, Bruckstein A. K-svd: An algorithm for designing of overcomplete dictionaries for sparse representation[J]. IEEE Transactions on Signal Process, 2006(11):4311—4322.

[18] Candès E J, Recht B. Exact matrix completion via convex optimization[J]. Foundations Of Computational Mathematics, 2009(9): 717—772.

[19] Ferguson, Thomas S. A Bayesian analysis of some nonparametric problems [J]. The Annals of Statist, 1973(1):209—230.

[20] Tipping M. Spare Bayesian learning and the relevance vector machine[J]. Journal of Machine Learning Research, 2001(1):211—244.

[21] Teh Y W, Jordan M I. Bayesian Nonparametrics: Hierarchical Bayesian Nonparametric Models with Applications [M]. Cambridge: Cambridge University Press, 2010.

[22] Sudderth, Erik B, Jordan, et al. Shared segmentation of natural scenes using dependent Pitman-Yor processes[C]//Paper presented at the 21th NIPS. Cambridge , MA,USA: MIT Press , 2008: 1585—1592.

[23] Peelenl M V, Fei—Fei L, Kastner S. Neural mechanisms of rapid natural scene categorization in human visual cortex[J]. Nature, 2009, 460 (7251): 94—97.

[24] Paisley J W, Zhou M, Sapiro G, et al. Nonparametric image interpolation and dictionary learning using spatially dependent Dirichlet and beta process priors[C]// ICIP, 2010: 1869—1872.

[25] Schummers J, Yu H, Sur M. Role of astrocytes in visual cortex: tuned responses, feature maps, and hemodynamic regulation[J]. Science, 2008 (320):1638—1643.

[26] Rozell C, Johnson D, Baraniuk R, et al. Sparse coding via thresholding and local competition in neural circuits[J]. Neural Computation, 2008, 20(10): 2526—2563.

[27] Bonnans J F, Gilbert J C, Lemarechal C, et al. Numerical optimization: theoretical and practical aspects[M]. New York: Springer, 2006.

[28] Borwein J M, Lewis A S. Convex analysis and nonlinear optimization: Theory and examples[M]. New York: Springer, 2006.

[29] d'Aspremont A, Bach F R, Ghaoui L E. Full regularization path for sparse principal component analysis[C]// Paper presented at the 24th International Conference on Machine Learning(ICML),2007: 177—184.

[30] Zass R, Shashua A. Nonnegative sparse PCA[C]//Paper presented at the 19th NIPS. Cambridge , MA,USA : MIT Press, 2006: 1561—1568.

[31] Kreutz-Delgado K, Murray J F, Rao B D, et al. Dictionary learning algorithms for sparse representation[J]. Neural Computation (NECO), 2003, 15(2):349—396.

[32] Candès E J, Tao T. The power of convex relaxation: Near-optimal matrix completion[J]. IEEE Transactions on Information Theory, 2010,56(5): 2053—2080.

[33] Honorio J, Samaras D, Paragios N. Sparse and locally constant Gaussian graphical models[C]//Paper presented at the 19th NIPS. Cambridge , MA, USA : MIT Press, 2009: 745—753.

[34] Huang J, Zhang T, et al. Learning with structured sparsity[C]//Paper presented at the 26th International Conference on Machine Learning (ICML), 2009: 53.

[35] A. M. Bruckstein, David L D, Elad M. From sparse solutions of systems of equations to sparse modeling of signals and images[J]. SIAM Review, 2009, 51(1):34—81.

[36] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society, Series B, 2006, 68 (1):49—67.

[37] Borwein J M, Lewis A S. Convex Optimization[M]. Cambridge: Cambridge University Press, 2004.

[38] Bradley Efron, Trevor Hastie, Iain Johnstone, et al. Least angle regression [J]. Annals of Statistics, 2002(32):407—499.

[39] Tseng P, Yun S. A coordinate gradient descent method for nonsmooth separable minimization[J]. Math Program, 2008(117): 387—423.

[40] Jolliffe I T. Principal component analysis[M]. New York: Springer, 2002.

[41] Shen H, Huang J Z. Sparse principal component analysis via regularized low rank matrix approximation[J]. Journal of Multivariate Analysis, 2008(99): 1015—1034.

[42] Cai J F, Candès E J, Shen Z. A singular value thresholding algorithm for matrix completion [J]. SIAM Journal on Optimization, 2010 (20): 1956—1982.

[43] Hoyer P O, Dayan P. Nonnegative matrix factorization with sparseness constraints[J]. Journal of Machine Learning Research, 2004(5):1457—1469.

[44] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(2):210—227.

[45] Yang J, Wright J, Huang T S, et al. Image super-resolution via sparse representation[J]. IEEE Trans on Image Processing , 2010, 19(11):2861—2873.

[46] Mairal J, Sapiro G, Elad M. Learning multiscale sparse representations for image and video restoration[R]. Technical report, DTIC Document, 2007.

[47] Cevher V, Sankaranarayanan A, Duarte M F, et al. Compressive sensing for background subtraction[C]//in European Conf. Comp. Vision, 2008: 155—168.

[48] Dikmen M, Huang T S. Robust estimation of foreground in surveillance videos by sparse error estimation[J]. Pattern Recognition, 2008(12):1—4.

[49] Rao S R, Tron R, Vidal R, et al. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories [J]. Computer Vision and Pattern Recognition, IEEE Computer Society Conference, 2008:1—8.

[50] Elhamifar E, Vidal R. Sparse subspace clustering[J]. Computer Vision and Pattern Recognition, IEEE Computer Society Conference, 2009 (0): 2790—2797.

[51] Ji S, Yuan L, Li Y X, et al. Drosophila gene expression pattern annotation using sparse features and term-term interactions[C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA: ACM, 2009: 407—416.

[52] Zhihua Z, Guang D. Optimal scoring for unsupervised learning[C]//Paper presented at the 22th NIPS. Cambridge, MA, USA: MIT Press, 2009: 36.

[53] Freedman D A. On the asymptotic behavior of Bayes's estimates in the discrete case[J]. The annals of Statistics, 1963(3):1368—1385.

[54] Rasmussen C, Williams C. Exponential Families of Stochastic Processes [M]. Berlin: Springer, 1997.

[55] Stein M L. Interpolation of Spatial Data: Some Theory for Kriging[M]. Berlin: Springer, 1999.

[56] Neal R M. Bayesian learning for neural networks[M]. NewYork: Springer, 1996.

[57] Neal R M. Monte carlo implementation of Gaussian process modeis for Bayesian regression and classification [R]. Department of Statistics, University of Toronto, 1997.

[58] M. N. Gibbs. Bayesian Gaussian processes for regression and classification [D]. University of Cambridge, 1997.

[59] Rasmussen C E, Rasmussen C E. Evaluation of Gaussian processes and other methods for nonlinear regression[D]. University of Toronto, 1996.

[60] Sethuraman J. A constructive definition of Dirichlet priors[J]. Statistica Sinica, 1994(4): 639—650.

[61] Korwar R M, Hollander M. Contributions to the theory of Dirichlet processes[J]. The annals of Probability, 1973(1):705—711.

[62] Diaconis P, Freedman D. On the consistency of Bayes estimates[J]. The annals of Statistics, 1986(11):1—25.

[63] Muller P, Quintana F A. Nonparametric Bayesian data analysis[J]. Statistical Science, 2004, 19(1):95—111.

[64] Ferguson T S, Phadia E G. Bayesian nonparametric estimation based on censored data[J]. Annals of Statistics, 1979(7):163—186.

[65] Wolpert R L, Ickstadt K, Hansen M B. A nonparametric Bayesian approach to inverse problems (with discussion)[J]. Bayesian Statistics, 2003(1): 403—418.

[66] James L F, Lijoi A, Prünster I. Conjugacy as a distinctive feature of the Dirichlet process [J]. Scandinavian Journal of Statistics, 2005 (33): 105—120.

[67] Hewitt E, Savage L J. Symmetric measures on Cartesian products[J]. Transactions of the American Mathematical Society, 1955, 80 (2): 470—501.

[68] Orbanz P. Infinite Dimensional Exponential Families in the Cluster Analysis of Structured Data[D]. University of Bonn, 2008.

[69] Ghosal S. Bayesian Nonparametrics(chap: The Dirichlet process, related priors and posterior asymptotics)[M]. Berlin: Cambridge University press, 2010.

[70] Xing W, Bruce C. Lda-based document models for adhoc retrieval[C]// Proceedings of the 29th Annual International SIGIR Conference. Washington,USA: 2006: 178—185.

[71] Xing E P, Sohn K. Hidden markov Dirichlet process: Modeling genetic recombination in open ancestral space[J]. Bayesian Analysis, 2007, 2(3): 501—528.

[72] Fox E, Sudderth E, Jordan M I, et al. An hdp-hmm for systems with state persistence [C]// In Advances in Neural Information Processing Systems 2009.

[73] Sudderth E, Jordan M I. Shared segmentation of natural scenes using dependent Pitman-Yor processes[C]// In Advances in Neural Information Processing Systems, 2009.

[74] Sudderth E, Torralba A, Willsky A. Describing visual scenes using transformed objects and parts[J]. International Journal of Computer Vision, 2008, 7(1): 291—330.

[75] Zhou M, Chen H, Paisley J, et al. Nonparametric Bayesian dictionary learning for sparse image representations[G]// Proc. Neural Inf. Process. Syst, 2009: 1—9.

[76] Shao J. Methmatical Statistics[M]. Berlin: Springer, 2003.

[77] Berstein S. Elements of Statistics II: Inferential Statistics[M]. New York: McGraw-Hill, 1999.

[78] Wasserman L. All of Nonparametric Statistics[M]. Berlin: Springer, 2006.

[79] Paisley J. Machine Learning with Dirichlet and Beta Process Priors: Theory and Applications[D]. Duke University, 2010.

[80] Pitman J. Combinatorial stochastic processes[R]. Department of Statistics, University of California at Berkeley, Lecture notes for St. Flour Summer School, 2002.

[81] Hjort N L. Nonparametric Bayes estimators based on beta processes in models for life history data[J]. The Annals of Statistics, 1990, 18(3):

1259—1294.

[82] Thibaux R, Jordan M I. Hierarchical beta processes and the Indian buffet process[C]// In Proceedings of the International Workshop on Artifficial Intelligence and Statistics, 2007.

[83] Donoho D L, Elad M, Temlyakov V. Stable recovery of sparse overcomplete representations in the presence of noise [J]. IEEE Transactions on Information Theory, 2006, 52(1): 6—18.

[84] Candès E, Romberg J, Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information [J]. IEEE Transactions on Information Theory, 2006, 52(2):489—509.

[85] Papoulis A, Pilai S U. Probability, random variables and stochastic processes[M]. New York: McGraw-Hill, 2002.

[86] Antoniadis A, Fan J. Regularization of wavelets approximations[J]. Joumal of the American Statistical Association, 2001, 96(1):939—967.

[87] Tipping M E. Sparse Bayesian learning and the relevance vector machine[J]. The Journal of Machine Learning Research, 2001(1): 211—244.

[88] Figueiredo M. Adaptive sparseness for supervised learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003 (25): 1150—1159.

[89] Caron F, Doucet A. Sparse Bayesian nonparametric regression [C]// In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 2008.

[90] Park T, Casella G. The Bayesian Lasso [J]. Journal of the American Statistical Association, 2008, 103(482):681—686.

[91] Lin Q, Lin N. The Bayesian elastic net[J]. Bayesian Analysis, 2010, 5(1): 151—170.

[92] Breiman L, Friedman J H. Estimating optimal transformations for multiple regression and correlation [J]. Journal of the American Statistical Association, 1985, 80(391): 580—598.

[93] Mallat S. A wavelet tour of signal processing[M]. New York: Academic Press, 1999.

[94] Candès E J, Donoho D L. Curvelets: a surprisingly effective nonadaptive representation for objects with edges[J]. Nashville, 1999 (2):105—120.

[95] Gersho A. Vector Quantization and Signal Compression [M]. Boston:

Kluwer Academic Publishers, 1992.

[96] Figueiredo M, Bioucas-Dias J, Nowak R. Majorization minimization algorithms for wavelet based image restoration[J]. IEEE Trans. Image Process, 2007, 16(12):2980—2991.

[97] Dong W, Li X, Zhang L, et al. Sparsity based image denoising vis dictionary learning and structural clustering [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[98] Candès E, Romberg J, Tao T. Stable signal recovery from incomplete and inaccurate measurements [J]. Communications on Pure and Applied Mathematics, 2006(8):1207—1223.

[99] Bishop C. Pattern Recognition and Machine Learning[M]. Berlin: Springer, 2006.

[100] Zivkovic Z, Heijden F. Efficient adaptive density estimation per image pixel for the task of background subtraction[J]. Pattern Recognition Letters, 2006, 27(7):773—780.

[101] Elliot D L. Covariance Regularization in Mixture of Gaussians for High Dimensional Image Classification[D]. Colorado State University, 2009.

[102] Socher R, Maas A, Manning C D. Spectral chinese restaurant processes: Nonparametric clustering based on similarities[C]// Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2011.

[103] Nia V P. Fast High Dimensional Bayesian Classification and Clustering [D]. Georgetown University, 2009.

[104] Adams R P, Ghahramani Z, Jordan M I. Tree structured stick breaking for hierarchical data[C]// Advances in Neural Information Processing (NIPS) 23. Cambridge, MA, USA: MIT Press, 2010.

[105] Mauldin R D, Sudderth W D, Williams S C. Pólya trees and random distributions[J]. The Annals of Statistics, 1992, 20(3):1203—1221.

[106] Lavine M. Some aspects of pólya tree distributions for statistical modelling [J]. The Annals of Statistics, 1992, 20(3):1222—1235.

[General Information]
□□=□□□□□□□□□□□□□□□□□□□□□□□□
□□=□□□
□□=90
SS□=13518997
DX□=
□□□□=2014.04
□□□=□□□□□□□□□